UNIVERSITÉ LIBRE DE BRUXELLES
FACULTÉ DE LETTRES, TRADUCTION ET COMMUNICATION

La gestion des données d'autorité archivistiques dans le cadre du Web de données

Anne CHARDONNENS

Thèse présentée en vue de l'obtention du grade académique de Docteure en Information et Communication sous la direction de Monsieur le Professeur Seth VAN HOOLAND

Année académique 2020-2021

Remerciements

Tout d'abord, je tiens à remercier sincèrement mon directeur de thèse, Seth van Hooland, pour la confiance qu'il m'a accordée depuis mes années de master déjà et pour l'autonomie qu'il m'a laissée, tout en me faisant profiter de ses conseils et connaissances. Ses publications, son sens de la formule qui fait mouche et son pragmatisme ont été une réelle inspiration pour moi. Je remercie également les membres de mon comité d'accompagnement pour leurs conseils éclairés, Isabelle Boydens, Françoise D'Hautcourt, Pierre-Alain Tallier et Sébastien de Valeriola.

Ma gratitude va également aux quatre personnes me faisant l'honneur de siéger dans mon jury de thèse et d'accorder temps et attention à ce travail : Florence Clavaud, Dirk Luyten, Max De Wilde et Pierre-Alain Tallier.

Sans être officiellement co-directrice de thèse sur papier, Florence Gillet, responsable de l'accès numérique aux collections au CegeSoma, l'a certainement été dans les faits. Bénéficier de son expertise, de son regard avisé et de son soutien enthousiaste a représenté une réelle plus-value au cours de cette expérience de recherche. Je retiendrai également son talent à encadrer les projets de recherche MADDLAIN et ADOCHS, dont les réunions ont eu le mérite d'animer et de rythmer un travail traditionnellement solitaire. Que soient également remerciés les promoteurs, les comités de suivi et les chercheurs de ces projets, en particulier Jill Hungenaert, regrettée voisine de bureau, et Stéphanie Paul, pour sa sagesse.

Ma thèse [1] s'est déroulée à cheval entre deux institutions, deux bureaux, deux équipes de collègues. Au-delà des défis logistiques que cela a pu représenter, je suis extrêmement reconnaissante d'avoir pu bénéficier de cette expérience enrichissante. Merci au CegeSoma et à tous ses collaborateurs qui m'ont accueillie avec chaleur humaine et bienveillance, me permettant d'affûter mon regard universitaire au contact du terrain et des discussions passionnantes qui n'ont pas manqué. Un merci particulier à Gertjan Desmet, Francesco Gelati, Dirk Luyten, Fabrice Maerten et Mathieu Roeges pour leur

1. Qui a pu être menée grâce au financement de la Politique scientifique belge dans le cadre des numéros de contrat BR/154/A5/MADDLAIN et BR/154/A6/ADOCHS, ainsi que grâce à un *Fellowship* de l'*European Holocaust Research Infrastructure* (EHRI).

disponibilité, ainsi qu'aux chercheurs sur projet pour nos séances de *Shut up and write* judicieusement orchestrées par Karla Vanraepenbusch. Je remercie également Yves Lardinois et Peter Van Overveldt, du service ICT des Archives de l'Etat, pour leur précieux soutien technique, ainsi que Ellen Van Keer, pour ses réponses à mes questions relatives au respect de la vie privée.

Merci également à mes (anciens) collègues doctorants de STIC, qui m'ont ouvert les portes d'un monde jusque-là inconnu, me permettant d'apprendre énormément à leur contact et d'apprivoiser une logique nouvelle. Mathias Coeckelbergs, Laurence Dierickx, Simon Hengchen, Raphaël Hubain, Laurence Maroye, Ettore Rizza : je ne serais pas arrivée là sans vous !

L'automne 2018 fut l'occasion d'un séjour au CDEC à Milan, dans le cadre d'un fellowship EHRI. L'occasion de préciser ma pensée et de prendre du recul grâce à Laura Brazzo, que je remercie pour son accueil chaleureux.

Thèse s'ancrant dans le contexte du web oblige, j'ai bénéficié de nombreuses aides et discussions en ligne, en particulier grâce à Twitter. Plus de dix ans après m'être inscrite sur ce réseau social, il est temps de lui rendre hommage et de souligner l'impact qu'il a eu sur mon parcours : inspirée par les #museogeeks, Museomix et les *Digital Humanities*, j'ai décidé de réaliser un mémoire sur le *crowdsourcing* et de poursuivre avec un doctorat, après avoir pu observer les coulisses de la thèse grâce aux tweets de la #teamrédaction. Puis, tout au long de ma recherche, j'ai tiré parti de la veille d'utilisateurs mettant si bien en œuvre l'adage *sharing is caring*. Merci notamment à Baptiste de Coulon, Sébastien Beyou, Olaf Simons, Magnus Salgö et Andra Waagmeester de m'avoir fait profiter de leur expertise.

Je ne pourrais pas oublier les précieux relecteurs et relectrices, qui, en plus de tout le reste, ont contribué à rendre cette thèse un peu plus lisible et digeste – je l'espère ! Merci à vous, Michelle, Florence, Baptiste et Mathias.

Enfin, je pense que tous les remerciements de thèse le prouvent, derrière chaque doctorant il y a une constellation de personnes qui sont là pour maintenir de la clarté dans les moments de tourmente, mais aussi pour partager les joies dans les moments d'épiphanie. J'aimerais exprimer en particulier ma gratitude à Cécile, Tali et Olivier, qui m'ont permis de tenir mon cap ; à Alexandra, Amélie, Aude, Julien, Romana, Salomé, Solange et Victory pour les séances de travail collectif à distance et le soutien indéfectible ; à ma famille et à ma belle-famille pour leur affection et leur patience ; à Orion, formidable compagnon à quatre pattes, pour sa présence fidèle et réconfortante au cours de mes longues nuits de travail, et enfin, à Julien, qui a eu la folie de m'encourager à me lancer dans cette périlleuse expédition et n'a eu de cesse de croire en moi.

Résumé

Dans un contexte archivistique en transition, marqué par l'évolution des normes internationales de description archivistique et le passage vers une logique de graphes d'entités, cette thèse se concentre plus spécifiquement sur la gestion des données d'autorité relatives à des personnes physiques. Elle vise à explorer comment le secteur des archives peut bénéficier du développement du Web de données pour favoriser une gestion soutenable de ses données d'autorité : de leur création à leur mise à disposition, en passant par leur maintenance et leur interconnexion avec d'autres ressources.

La première partie de la thèse est dédiée à un état de l'art englobant tant les récentes évolutions des normes internationales de description archivistique que le développement de l'écosystème Wikibase. La seconde partie vise à analyser les possibilités et les limites d'une approche faisant appel au logiciel libre Wikibase. Cette seconde partie s'appuie sur une étude empirique menée dans le contexte du Centre d'Études et de Documentation Guerre et Sociétés Contemporaines (CegeSoma). Elle permet de tester les perspectives dont disposent des institutions possédant des ressources limitées et n'ayant pas encore adopté la logique du Web de données. Par le biais de jeux de données relatifs à des personnes liées à la Seconde Guerre mondiale, elle dissèque les différentes étapes conduisant à leur publication sous forme de données ouvertes et liées.

L'expérience menée en seconde partie de thèse montre comment une base de connaissance mue par un logiciel tel que Wikibase rationalise la création de données d'autorité structurées multilingues. Des exemples illustrent la façon dont ces entités peuvent ensuite être réutilisées et enrichies à l'aide de données externes dans le cadre d'interfaces destinées au grand public. Tout en soulignant les limites propres à l'utilisation de Wikibase, cette thèse met en lumière ses possibilités, en particulier dans le cadre de la maintenance des données.

Grâce à son caractère empirique et aux recommandations qu'elle formule, cette recherche contribue ainsi aux efforts et réflexions menés dans le cadre de la transition des métadonnées archivistiques.

Abstract

The subject of this thesis is the management of authority records for persons. The research was conducted in an archival context in transition, which was marked by the evolution of international standards of archival description and a shift towards the application of knowledge graphs. The aim of this thesis is to explore how the archival sector can benefit from the developments concerning Linked Data in order to ensure the sustainable management of authority records. Attention is not only devoted to the creation of the records and how they are made available but also to their maintenance and their interlinking with other resources.

The first part of this thesis addresses the state of the art of the developments concerning the international standards of archival description as well as those regarding the Wikibase ecosystem. The second part presents an analysis of the possibilities and limits associated with an approach in which the free software Wikibase is used. The analysis is based on an empirical study carried out with data of the Study and Documentation Centre War and Contemporary Society (CegeSoma). It explores the options that are available to institutions that have limited resources and that have not yet implemented Linked Data. Datasets that contain information of people linked to the Second World War were used to examine the different stages involved in the publication of data as Linked Open Data.

The experiment carried out in the second part of the thesis shows how a knowledge base driven by software such as Wikibase streamlines the creation of multilingual structured authority data. Examples illustrate how these entities can then be reused and enriched by using external data in interfaces aimed at the general public. This thesis highlights the possibilities of Wikibase, particularly in the context of data maintenance, without ignoring the limitations associated with its use.

Due to its empirical nature and the formulated recommendations, this thesis contributes to the efforts and reflections carried out within the framework of the transition of archival metadata.

Table des matières

Table des figures

Liste des tableaux

Préambule

Afin de faciliter la lecture de cette thèse, qui comprend des citations en langue étrangère, mais également des éléments plus techniques comme des extraits de code informatique, des éléments issus de la base de connaissance en ligne Wikidata et de questions formulées selon le langage de requête SPARQL, plusieurs choix ont dû être posés. Nous les détaillons ici :

1. Les membres de notre jury étant, à notre connaissance, capables de lire l'anglais, l'ensemble des citations en anglais n'a fait l'objet ni de traductions, ni d'une mise en évidence particulière – afin d'éviter une surchage visuelle.
2. Lorsque c'était nécessaire, certaines citations ont été légèrement reformatées afin de préserver un usage uniforme des guillemets, points de suspension, mentions de siècles, tirets d'incise et autres conventions typographiques. Cette harmonisation a été réalisée avec prudence et minutie, en veillant à ne pas dénaturer les propos de l'auteur. En revanche, les usages de ponctuation hasardeuse ou les fautes d'orthographe – rencontrés par exemple parmi des sources plus informelles comme des tweets – ont été préservés, dans un souci d'authenticité.
3. Les références à des entités issues de la base de connaissance Wikidata ou d'une autre instance issue du logiciel Wikibase sont signalées ainsi : « Q31|Belgique »[2], à savoir l'identifiant numérique de l'élément accompagné du libellé en français, suivi d'un appel de note de bas de page renseignant le lien pour accéder à la page web de l'entité concernée.
4. Certaines figures sont issues de documents ayant été publiés sous licence *Creative Commons*, nous le signalons dans la mention de la source à l'aide de l'abréviation adéquate (comme par exemple *CC BY-SA*). Le détail sur les conditions de chacune de ces licences peut être consulté sur le site Creative Commons[3].

2. https://www.wikidata.org/wiki/Q31
3. https://creativecommons.org/.

5. Les propos d'utilisateurs diffusés sous forme de tweets font l'objet d'une mise en page distincte (bloc de couleur grise encadré de guillemets de grande taille). Ces tweets cités étant susceptibles d'être supprimés, leur mention dans la bibliographie a été accompagnée d'un lien Internet Archive vers une version archivée garantissant leur accessibilité.
6. Les extraits de conversations écrites issues de groupes Telegram ne pouvant être cités de façon pérenne uniquement s'il s'agit d'un groupe public, une parade a dû être trouvée [4]. Faute de mieux, nous avons donc dû nous contenter d'accompagner, par le biais d'une note de bas de page, les extraits cités de la mention [Message Telegram], suivie du nom de l'utilisateur et de la date d'émission du message.
7. Les diverses pages Wikimédia [5] citées dans cette thèse étant co-rédigées par un nombre parfois important de contributeurs, nous avons pris le parti de ne pas inclure une liste exhaustive de ces auteurs et de privilégier une attribution générique (comme par exemple *Wikidata*), à moins qu'il ne s'agisse d'un cas particulier, comme la page personnelle d'une utilisatrice ou d'un utilisateur.
8. L'utilisation de termes épicènes a été privilégiée dans cette thèse. Dans les cas où c'est le genre masculin qui a été utilisé, il doit évidemment être entendu comme inclusif et incluant tous les genres.
9. Enfin, étant donné le caractère empirique de la seconde partie de cette thèse, basée sur la manipulation et la transformation de données, sur l'usage d'interfaces en ligne et de scripts de programmation, nous avons décidé, sur suggestion de notre directeur de thèse, d'associer à cette thèse des *annexes interactives*, en ligne [6]. Ces dernières, signalées à l'aide de notes de bas de page aux endroits correspondants, doivent permettre, d'une part, d'alléger la lecture en ne surchargeant pas inutilement le texte de détails techniques, d'autre part, de pallier les limites du format papier [7] et d'illustrer plus facilement certaines étapes-clé de cette thèse.

4. En effet, le groupe Telegram *Wikibase Community*, fréquemment mentionné dans cette thèse, a beau être accessible à tous à l'aide d'un lien (de même que son historique), il n'en reste pas moins privé.
5. Nous englobons ici l'ensemble des pages associées aux projets Wikimédia, qu'il s'agisse de Wikidata, de Wikipédia en français ou encore de MediaWiki.
6. https://linkingthepast.org.
7. Nous les avons toutefois également intégrées sous format papier, par souci de complétude et practicité pour les lecteurs.

Introduction

Contexte

> Ôter son nom à un être humain et social, c'est lui ôter quelque chose de sa personne. Mais je conviens que le vocable de Radégonde est rude.
>
> *Anatole France*

En décembre 2018, quiconque voulant effectuer une recherche sur Andrée de Jongh, la co-fondatrice du réseau de résistance belge Comète, devait attendre la deuxième page des résultats de Google pour trouver un lien renvoyant vers les ressources en ligne du CegeSoma (le Centre d'Études et de Documentation Guerre et Sociétés contemporaines). Ce lien renvoyait vers une biographie mise à disposition en PDF [8] sur le site du CegeSoma. Depuis, Belgium WWII, la plateforme consacrée à la Belgique durant la Seconde Guerre mondiale et mise en ligne par le CegeSoma en septembre 2017, a dédié une page *Personnalités* à Andrée de Jongh [9], qui apparaît désormais en seconde place des résultats de recherche Google.

Cependant, qu'il s'agisse de la biographie diffusée au format PDF ou de la page dédiée de Belgium WWII, ces informations existent de manière isolée. En effet, la personne désireuse de trouver des documents d'archives associés à cette figure de la résistance belge devra effectuer encore d'autres recherches dans le catalogue du CegeSoma. Grâce à une recherche par mot-clé, elle pourra trouver les documents associés à l'entrée du *thésaurus* [10] du Centre *de jongh, andree (alias dedee / postman ; 1916-2007)* : des archives [11],

8. http://www.cegesoma.be/docs/images/stories/ceges/Accueil/Andr__e_De_Jongh___EN.pdf.

9. https://www.belgiumwwii.be/belgique-en-guerre/personnalites/andree-de-jongh.html.

10. Comme nous le verrons au cours de la seconde partie de cette thèse, cette appellation – thésaurus – est discutable, nous faisons toutefois le choix d'utiliser ce terme *thésaurus* afin de rester fidèle à la dénomination officielle de cette liste (voir Temmermann et Waeyenberg, 2000).

11. Le CegeSoma a récemment acquis un fond d'archives privées de Andrée de Jongh ; lorsque ce dernier aura été inventorié, il y aura potentiellement une nouvelle autorité Per-

des livres ou articles, une coupure de presse et des photos. Cependant, elle devra persévérer et investiguer plus avant afin de voir si les documents associés aux mots-clés *de jongh* ; *de jonghe* ; *de jonghe, a.* la concernent également. Par ailleurs, elle découvrira que plusieurs informations essentielles se trouvent encodées exclusivement dans la description associée aux coupures de presse :

> [Coupures de presse concernant] Andrée DE JONGH, dite Dédé, alias Postman, résistante belge, une des fondatrices du réseau d'évasion Comète, arrêtée en 1943 et déportée à Ravensbrück et Mauthausen (1916-2007).

Bien que ces diverses informations soient pour le moment éclatées, elles ont le mérite de toutes se trouver mises en ligne. En revanche, il faut consulter un document Excel – disponible en interne uniquement – pour voir que son nom figure également sur la liste des agents du Service de Renseignements et d'Action, auxquels sont associés des dossiers individuels conservés par le CegeSoma [12].

Cet exemple illustre la façon dont l'information est disponible de façon fragmentaire au sein même de l'institution. Cependant, cela ne s'arrête pas là. Le même phénomène se produit à une échelle supérieure : rien ne relie actuellement le mot-clé du CegeSoma *de jongh, andree (alias dedee / postman ; 1916-2007)* à d'autres ressources la concernant, conservées dans diverses institutions à travers le monde. Or, son parcours de résistante a conduit Andrée de Jongh à effectuer des activités de résistance hors du territoire belge : elle a aidé nombre d'aviateurs anglais à rejoindre leur pays en passant par la France et l'Espagne, avant d'être arrêtée et de passer par diverses prisons et camps de concentration allemands. Quelques recherches nous révèlent ainsi que des institutions comme l'Imperial War Museum, the National Archives, the Library of Congress, la Bibliothèque nationale de France, la Bibliothèque nationale des Pays-Bas ou encore le Ministère français de la Défense (Mémoire des Hommes) possèdent des ressources relatives à Andrée de Jongh, comme l'illustre la figure 1.

Outre le caractère laborieux de cette tâche nécessitant de consulter différents catalogues en ligne sans aucune assurance de résultats, il faut également garder à l'esprit qu'une recherche à l'aide de la chaîne de caractères *Andrée de Jongh* ne garantit pas de retrouver tous les résultats pertinents : on la retrouve par exemple sous la forme retenue de *De Jongh, Andrée Eugé-*

sonne (auteurs et producteurs) qui sera créée pour Andrée de Jongh, indépendante du mot-clé précité. C'est le cas par exemple pour Léo Lejeune, qui est à la fois producteur et sujet des collections du CegeSoma.

12. Dossiers Services de Renseignement et d'Action SRA de la Sûreté de l'État, Fonds AA1333.

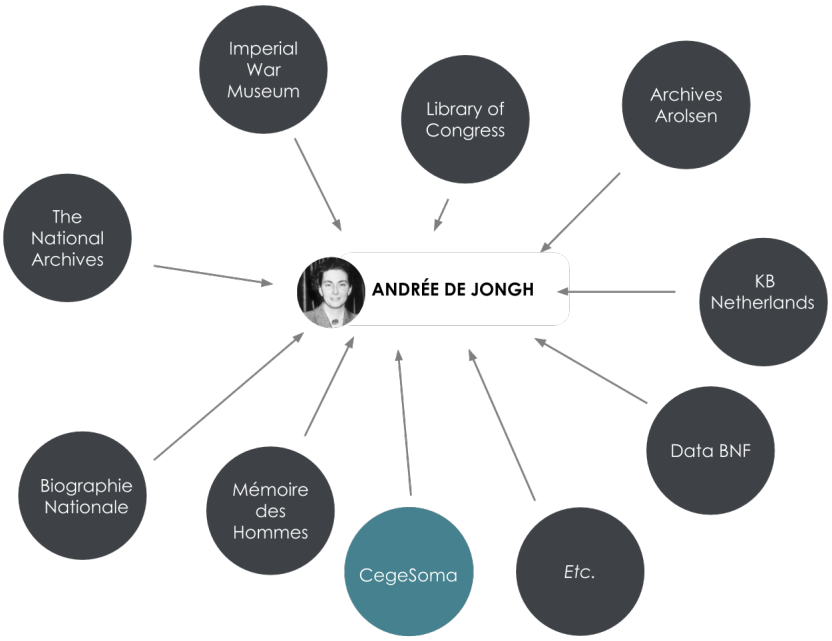


FIGURE 1 – Outre le CegeSoma, plusieurs institutions possèdent des ressources sur la résistante belge Andrée de Jongh.

nie Adrienne (1916-2007) ou encore *Countess Andree « Dedee » de Jongh*. En poursuivant l'exploration, on découvre qu'elle était également connue sous d'autres noms de code comme *Cyclone DD*, *Petit Cyclone*, *Little Cyclone*. Si un être humain est capable de faire le lien entre ces différentes appellations grâce à son esprit de déduction, certaines équivalences sont en revanche difficilement compréhensibles par un ordinateur si elles n'existent que sous forme de chaînes de caractères en langage naturel. En effet, là où un algorithme de *fuzzy matching* pourra potentiellement établir un lien entre deux chaînes de caractères proches telles que *andre de jong* et *Andrée de Jongh*, il ne pourra pas le faire entre *Andrée de Jongh* et *Petit Cyclone*, à moins qu'un lien d'équivalence n'ait été établi dans un format compréhensible par une machine.

Et encore, il s'agit là d'un prénom et d'un nom de famille comportant certes une particule, mais pouvant être assez aisément identifiés comme tels, ce qui est loin d'être toujours le cas. En effet, les noms de personnes sont sujets à divers problèmes d'ambiguïté. La presse nous donne un bon aperçu de cette complexité. Ainsi, lorsque ce ne sont pas des noms d'aliments qui sont utilisés comme prénoms [13], ce sont des noms de villes en hommage à

13. Les statistiques du site BabyCenter concernant les prénoms les plus populaires aux États-Unis pour l'année 2018 révèlent une tendance de plus en plus marquée à choisir comme prénom des noms d'aliments sains tels que Kiwi, Kale ou encore Saffron (Barthélémy, 2018).

une bataille historique [14]. Outre l'ambiguïté née de ces cas d'homonymie, de particularités locales [15] ou de multiples variations orthographiques pouvant concerner un même nom [16], il faut également prendre en considération les titres et suffixes [17], les pseudonymes [18] ou encore les changements de noms [19]. Sans parler des noms issus d'autres époques [20] ou d'autres cultures, impliquant des translittérations, des caractères spéciaux, des longueurs ou structures atypiques pour un œil habitué aux noms occidentaux ou encore d'autres types de particularités, qui peuvent avoir de nombreuses implications dans le cadre du Web [21].

Dans ce monde de complexité, où des milliers d'inconnus se partagent le vocable de Jane ou John Doe [22], les institutions proposant un accès numérique à leurs collections, comme les bibliothèques, les archives et les musées, ne peuvent donc pas se reposer uniquement sur de simples chaînes de caractères. Elles ont besoin d'autorités, c'est-à-dire de notices de référence, qui constituent en quelque sorte la carte d'identité de l'entité et permettent de l'identifier de manière non équivoque. Ces notices servent à la fois à gérer, à normaliser et à garantir un accès cohérent à l'information (Bourdon, 1997).

14. Ainsi, une étude décrite sur le blog tenu par The National Archives relève que 1 634 prénoms liés à la Première Guerre Mondiale ont été donnés à des bébés nés en Angleterre et au Pays de Galles entre 1914 et 1939, selon les *General Register Office indexes*. Il s'agit notamment de noms inspirés de célèbres batailles, les prénoms les plus populaires étant Verdun (901 prénoms), Ypres (71) et Mons (58) (Jessamy, 2016).

15. Comme le système de noms islandais, qui identifie la plupart des individus à l'aide de patronymes ou matronymes, auxquels s'est ajouté en 2019 un suffixe neutre – bur –, dans le cadre d'une modification de la loi sur les personnes visant à inclure les personnes non-binaires (Ragnarsdóttir, 2019).

16. Nous pouvons par exemple penser au cas d'école de Moammar Gaddafi, dont plus de 110 graphies anglaises différentes ont été recensées (Bass, 2009).

17. Junior – abrégé Jr –, est par exemple indispensable pour distinguer un fils de son père dans le contexte d'écrivains américains possédant le même anthroponyme et dont les dates biographiques ne seraient pas connues (Krieger et ABES, 2016, p. 15-16).

18. C'est le cas par exemple de l'écrivain belge Louis Carette, plus connu sous son nom de plume, Félicien Marceau.

19. Par exemple dans le cadre d'un mariage, mais pas seulement : ainsi, en Finlande, conformément à la loi de 2017 sur les noms et prénoms, n'importe quel adulte résident de manière permanente dans le pays a la possibilité de changer de nom, sans justification, à l'aide d'une simple demande en ligne (Digital and Population Data Services Agency, 2020).

20. Voir par exemple les difficultés auxquelles doit faire face Trismegistos, une base de données de ressources épigraphiques et papyrologiques : https://www.trismegistos.org/ref/about_namvars.php.

21. Comme le soulignent les recommandations du W3C (W3C et Ishida, 2011), cette liste énumérant diverses fausses croyances de programmeurs au sujet des noms de personnes (McKenzie, 2010), ou encore cet article de linguistique computationnelle portant les difficultés propres à la recherche de noms sur Wikipedia (Jones, 2018).

22. Ce nom est traditionnellement assigné à des personnes non identifiées dans le cadre de poursuites judiciaires ou de décès. Ainsi, Paulozzi *et al.* ont recensé aux États-Unis, pour la période allant de 1979 à 2004, une moyenne annuelle de 413 personnes décédées auxquelles ce nom a été attribué (Paulozzi *et al.*, 2008).

Ces notices d'autorité, dont l'apparition standardisée dans le secteur des archives remonte à l'année 1995 avec le lancement de la première édition de la norme ISAAR (CPF) [23], visant à séparer la description des archives de leur producteur, sont aujourd'hui appelées à évoluer. Les archives commencent en effet, à l'instar des bibliothèques [24], à préparer leur transition [25]. Or, comme le montrent les journées d'études consacrées à ce sujet, cela commence par une attention particulière portée aux données d'autorité, « l'or noir » des [archives et] bibliothèques :

> N'oubliez pas les données d'autorité ! Pourquoi et comment les choyer pour préparer les futures entités de votre catalogue. (ABES et BNF, 2019a)

Cette transition implique notamment, dans le cadre des archives, de passer d'une logique de notices d'autorité centrées sur les producteurs d'archives [26] à la généralisation de leur utilisation également pour la description du contenu des documents. En effet, il s'agit d'un usage étendu de l'indexation des personnes dans tous les aspects des descriptions d'archives, en utilisant pour cela les mêmes listes d'autorités, qui sont désormais envisagées sous la forme d'entités (Archives nationales de France, 2019). Ce changement paradigmatique est favorisé par l'évolution des technologies, en particulier de la technologie des données liées et du Web de données que nous allons explorer dans cette thèse.

Cette transition doit répondre à différents enjeux, d'ordre tant technique – de par la forme de l'encodage des informations – que conceptuel – avec l'élaboration de nouveaux modèles intellectuels de description et les choix éditoriaux liés à la forme de publication de ces informations. En effet, les notices d'autorité créés par les services d'archives s'inscrivant désormais dans le contexte du Web de données, elles doivent pouvoir être publiées sous la forme de jeux de données en RDF [27] à même d'être interrogés à l'aide du langage de requête SPARQL (International Council on Archives Expert

23. International Standard Archival Authority Record for Corporate Bodies, Persons and Families (International Council on Archives, 1996b)

24. Voir : https://www.transition-bibliographique.fr/.

25. Les Archives nationales de France organisent ainsi le 28 janvier 2020 une première journée dédiée aux « métadonnées archivistiques en transition », annonçant qu'à la suite des bibliothèques françaises, « les services d'archives entrent aujourd'hui dans ce processus » (Archives nationales de France, 2019).

26. La première édition de la norme ISAAR (CPF) se concentrait en effet sur les producteurs d'archives, bien qu'elle spécifie qu'« une notice d'autorité archivistique conforme à cette norme peut aussi servir à contrôler la forme du nom et l'identité de toute collectivité, personne ou famille citée dans tout point d'accès de la description » (International Council on Archives, 1996b).

27. RDF, pour *Resource Description Framework*, est le standard du Web sémantique pour la description de ressources. Pour davantage d'informations, voir les ressources du W3C, par exemple : https://www.w3.org/TR/rdf11-concepts/.

Group on Archival Description, 2019a) et d'inclure des liens d'équivalence entre entités, au-delà des murs d'une même institution ou d'un même pays. De plus, il s'agit également de pouvoir faire face à une complexité croissante, par exemple pour abriter des perspectives plurielles ou représenter des provenances multiples et complexes (Douglas *et al.*, 2018; Antracoli et Rawdon, 2019). Disposer d'outils de description plus performants et transparents doit ainsi permettre aux institutions des gains de temps sur le long terme, mais également de disposer de fonctionnalités plus fines permettant par exemple d'offrir un accès à l'historique des modifications (Pintscher *et al.*, 2019b, p. 11). Ainsi, une institution décidant de réviser des métadonnées problématiques contenant, par exemple, des termes racistes (Antracoli et Rawdon, 2019), devrait avoir la possibilité de le faire sans devoir renoncer au fait de pouvoir préserver une copie de la version précédente [28].

En outre, l'enjeu n'est plus seulement de pouvoir échanger de l'information contextuelle (International Council on Archives, 1996b), mais bien de pouvoir mutualiser sa production [29] et d'expérimenter des formes de production décentralisées de connaissances. D'autres enjeux sont plus directement liés aux utilisateurs finaux, qu'il s'agisse de fournir de nouveaux modes d'accès et de visualisation des données tirant parti de leur caractère structuré et de leur richesse sémantique, ou encore de modes de contribution collaboratifs. Enfin, il faut relever que l'évolution des fichiers d'autorité traditionnels vers des référentiels reposant sur une logique de graphe se traduit par un impact bidirectionnel : d'une part, les technologies du Web sémantique s'invitent au sein des archives, bibliothèques et musées ; d'autre part, les notices d'autorité de ces institutions s'invitent dans le Web de données : ces données sont en effet décrites comme ayant le potentiel de devenir « the backbone of a machine-readable, semantic web of culture and science » (Pintscher *et al.*, 2019b, p. 12). Le contexte étant maintenant posé, nous pouvons nous intéresser au cadre théorique qui va nous permettre d'affûter notre vision et d'appréhender cette situation de transition des données d'autorité archivistique à travers l'angle plus spécifique de la maintenance. Une fois ce cadre exposé, nous introduirons les questions de recherche autour desquelles est construite cette thèse.

28. Voir Drabinski (2013, pp. 108-109).

29. « While the library community has a long history of cooperatively describing library holdings and sharing the descriptive data, the archival community does not. [...] While the archival holdings are largely unique, they are nevertheless inextricably interconnected in ways that provide opportunities for cooperating that will significantly benefit both the archival community and the researchers, scholars, and members of the public that use their holdings. » (SNAC, 2017).

Cadre théorique et questions de recherche

> What happens when we take erosion, breakdown, and decay, rather than novelty, growth, and progress, as our starting points in thinking through the nature, use, and effects of information technology and new media
>
> Steven Jackson

Cette section détaille le cadre théorique dans lequel s'inscrit cette thèse et sur lequel reposent les questions de recherche qui seront introduites dans un deuxième temps. Les prochaines pages visent à mobiliser les concepts de maintenance, de *broken world thinking* et de dette technique pour éclairer la problématique de la gestion des données d'autorité dans un contexte archivistique en pleine évolution, qui sera décrit en détail au cours des prochains chapitres.

Russell et Vinsel définissent la maintenance comme « all of the work that goes into preserving technical and physical orders » (Russell et Vinsel, 2018, p. 18). Ce travail est réalisé par des *maintainers*, « those individuals whose work keeps ordinary existence going rather than introducing novel things » (Russell et Vinsel, 2016). Si cette notion suscite de plus en plus d'intérêt dans de nombreuses disciplines académiques et pratiques professionnelles [30], de l'architecture aux sciences de l'information (Mattern, 2018), le sujet n'est pas pour autant nouveau.

30. Il est d'ailleurs difficile de ne pas faire de lien avec l'actualité, ces lignes étant rédigées au moment où l'Europe se confine pour limiter la propagation du coronavirus Covid-19, les *maintainers* sortant soudainement de l'ombre : des livreurs aux réassortisseurs de rayons de pâtes alimentaires, en passant par le personnel soignant qu'applaudissent désormais chaque soir de leur balcon les confinés, sans oublier le personnel de nettoyage dont le travail est enfin valorisé.

Nous pouvons par exemple penser aux chercheurs de Palo Alto qui se sont intéressés dans les années 1980-1990 aux interactions entre humain et machine à travers le prisme de l'anthropologie. Ainsi, Lucy Schuman et Julian Orr, membres du centre de recherche Xerox, vont s'intéresser de près aux photocopieuses. Dans son ouvrage *Plans and Situated Actions, The problem of Human-machine Communication*, Schuman (1987) présente une étude de cas qui fera date : elle analyse l'interaction entre des utilisateurs inexpérimentés et une photocopieuse Xerox munie d'un système expert d'aide. Elle observe que de cet acte qui peut paraître simple à première vue naissent divers malentendus et elle en dissèque les causes, de l'incomplétude des instructions aux divergences liées aux termes utilisés en passant par la part d'implicite [31].

Son collègue Orr s'intéresse à l'étape qui suit l'utilisation quotidienne : les pannes et réparations des machines. Il va suivre le quotidien de techniciens spécialisés dans la réparation de photocopieuses dans une grande entreprise américaine. Dans l'introduction de son ouvrage *Talking about machines*, Orr affirme qu'étudier une pratique permet de révéler qu'un travail est généralement différent et souvent plus complexe qu'imaginé. Ainsi, sa présence sur le terrain va lui permettre de mettre en exergue tous les dysfonctionnements imprévisibles des machines, non pris en compte par les manuels et procédures officielles des organisations. Il relève qu'au-delà des réparations de routine, de nouveaux types de défaillance apparaissent sans cesse, que les procédures ne peuvent pas anticiper et qu'elles devront être prises en charge par les techniciens :

> Technicians' practice is therefore a response to the fragility of available understandings of the problematic situations of service and to the fragility of control over their definition and resolution. (Orr, 2006, p. 2)

En effet, des informations précises sur l'état d'une machine ne sont pas toujours disponibles et quand elles le sont, leur signification n'est pas toujours claire. Par ailleurs, le contrôle est fragile dans la mesure où le technicien est sollicité lorsque la panne s'est déjà produite et que la relation entre le client et la machine est déjà mise à mal. « Work in such circumstances is resistant to rationalization, since the expertise vital to such contingent and extemporaneous practice cannot be easily codified », affirme ainsi Orr (1996, pp. 1-2).

Grâce à sa posture d'observateur, Orr peut se faire le témoin de la dimension collective que prend la maintenance et la réparation des photocopieuses, « a continuous, highly skilled improvisation within a triangular re-

31. Par exemple : une instruction donnée une seule fois et sous-entendue par la suite.

lationship of technician, customer, and machine » (Orr, 1996, p. 1), mettant en exergue l'importance des dialogues informels entre techniciens [32].

La publication de cette étude ayant suscité de nombreuses réactions, Orr y réagit dix ans plus tard avec son article *Ten years of talking about machines*. Dans ce contexte, il met en exergue un principe-clé :

> It is true that not all repair job are well done, but the need to repair a previously repaired machine does not necessarily suggest incompetence. At this point, the fact that I was a technician should qualify anyone's reading of what follows. Among people who work with machines, failure is expected. Machine in use wear out and break ; it is thought normal then to repair them, but it is not seen as permanent. It is expected that they will fail again with further use. (Orr, 2006, pp. 1810-1811)

C'est cette même réalité qu'observent de nombreux chercheurs, à l'instar des sociologues Jérôme Denis et David Pontille dans le cadre de leur étude sur la signalétique du métro parisien :

> For those who repair and monitor them, signs are never truly stable. Their colors fade, they wear out, their surface is attacked by mold, they are stolen, they break [. . .] Awareness of the instability of the signboards and the changes they undergo is at the heart of these workers' job. It is the essence of their expertise. (Denis et Pontille, 2014, p. 412)

Ayant accompagné les équipes de maintenance des stations de métro, les sociologues admettent avoir découvert un monde moins stable qu'imaginé :

> From day to day, the supervision and repair operations ensure the sturdiness, the sustainability and the efficiency of the signs network intended to guide the travelers. (CSI Mines ParisTech, 2016)

En prenant appui sur l'ouvrage de Bruno Latour *Reassembling the Social*, l'un des membres du réseau de recherche *The Maintainers* [33] attire l'attention sur ce même concept de *decay as a status quo*, soulignant qu'une fois que cette notion de détérioration continue est acceptée, il devient évident que la maintenance n'est pas seulement nécessaire lorsque les choses cessent de fonctionner : « rather, the mundane activities of maintenance are what make most things function to begin with – it is the very stuff our society is made of » (Kolkman, 2016).

32. Ces derniers sont en effet ponctués de *war stories* qui s'avèrent cruciales pour la pose des diagnostiques et l'identification de ce qui ne fonctionne pas.

33. https://themaintainers.org/.

C'est également ce qu'affirme sans ambages Mattern (2018) : « infrastructures fail everywhere, all the time. [...] Now breakdown is our epistemic and experiential reality. » Dans son article intitulé *Maintenance and Care*, elle explique qu'il est dès lors nécessaire d'étudier « how the world gets put back together. », c'est-à-dire le travail quotidien d'entretien, de maintenance, de soin et de réparation. En attendant que la maintenance puisse défier l'innovation en tant que paradigme dominant, et au vu du « degree of brokenness of the broken world », elle propose dans cet essai d'envisager la maintenance comme cadre correctif, en commençant par étudier les pratiques de maintenance préexistantes. Elle discerne quatre échelles de maintenance, de *Rust* (réparations de larges infrastructures urbaines) à *Corruption*, qui nous intéresse particulièrement dans la mesure où cette échelle concerne le nettoyage et la maintenance des données [34].

Avec l'échelle intitulée *Corruption*, Mattern embrasse d'un même regard toutes les activités de maintenance relatives au code informatique et aux données. Elle explique qu'à l'instar des villes et des bâtiments, la plupart des infrastructures logicielles tomberaient rapidement en panne sans l'intervention d'agents de maintenance.

Au sein des exemples cités par l'auteure sous le regroupement de *Corruption*, différentes échelles pourraient également être distinguées, des grands projets de cyber-infrastructures aux jeux de données diffusés à des fins de recherche ou de marketing, en passant par les logiciels libres, mais l'importance de la maintenance et sa tendance à être invisibilisée restent la même [35]. Elle illustre cette tendance à l'invisibilisation en citant une étude de cas particulièrement percutante dans le cadre de cette thèse : celle de l'ethnographe Jean-Christophe Plantin qui s'est intéressé au travail de préparation des jeux de données dans le contexte des sciences sociales et qui relève qu'un jeu de données « must look pristine at the end of its processing ». En effet, bien que de multiples interventions soient nécessaires avant que les données puissent être réutilisées [36], ce travail de maintenance n'est

34. Les deux autres échelles décrites par Mattern sont *Dust*, qui comprend l'entretien architectural mais également les travaux ménagers ou toute autre forme d'entretien domestique, et *Cracks*, qui englobe les activités de réparations d'objets, du téléviseur aux téléphones portables.

35. Comme elle le souligne, en reprenant les termes du premier *Festival of Maintenance* : « the maintenance of open-source software, online communities, co-ops, and datasets is not unlike the maintenance of natural environments, infrastructures, industries, cultural heritages, and material resources » (https://festivalofmaintenance.org.uk/, cité par Mattern, 2018).

36. À ce sujet, voir également les travaux de Denis et Goëta qui ont examiné ces interventions dans le cadre de l'open data : « les données brutes de l'open data sont le résultat d'opérations qui visent à désencastrer les données de ces réseaux, à les débarasser de la ganguge des pratiques qui en faisaient des données de métier, pour les transformer en données ambiguës, *ouvertes* à de nombreux types de traitements. Autrement dit, les données sont

pas sensé être visible pour les utilisateurs finaux qui ont tendance à croire qu'ils vont travailler avec des données *brutes* (Plantin, 2018, cité par Mattern, 2018).

Cette notion d'invisibilisation est également présente dans les recherches menées par Downey sur les nouveaux médias. En mettant en parallèle les tendances actuelles avec les pratiques médiatiques du XIX^e^ siècle, cet historien met en lumière le rôle essentiel de l'*human information labor*, qui contribue à la circulation de l'information, qu'il s'agisse de données, de contenu ou de connaissances. Il invite à se demander, tant dans le contexte de l'industrie télégraphique au XIX^e^ siècle que dans celui de plateformes telles que Google, Wikipedia, Facebook ou Amazon aujourd'hui : « who does what kind of information work, when and where and why ? » En se penchant sur cette question, Downey a observé que

> All of these forms of information labor share a crucial aspect, however : users tend not to see it. [...] Information laborers of all sorts are likely to be hidden, out of sight and out of mind, from those who encounter their products and processes on a daily basis. (Downey, 2014, p. 145)

Ses recherches l'ont également amené à s'intéresser aux bibliotechniciens[37] du milieu du XX^e^ siècle : ces derniers, impliqués dans des tâches telles que l'acquisition, le catalogage, la classification, l'indexation, la recherche et l'extraction d'informations, font face à l'informatisation du secteur. Ils doivent s'adapter à une progression « longue et coûteuse » des catalogues papier isolés vers des systèmes distribués, qui va entraîner une modification de leurs tâches, mais aussi de leur statut : de gardiens du savoir, ils deviennent des « librarians as information analysts and consultants ». S'ils prennent part à un processus de transcodage, en transformant des demandes d'information en langage de requête structurée pour interroger des bases de données ou encore en interprétant des listes de résultats devant être expliqués à des usagers de la bibliothèque, leur rôle ne s'arrête pas là. Il comporte également une dimension de production intellectuelle, consistant en un travail de contextualisation des documents à travers la création de « méta-information »[38]. Ce travail de contextualisation est nécessaire pour que l'information puisse circuler d'un contexte à l'autre, que ce soit à travers

brutes lorsque l'on réussit à les dé-spécifier de leurs usages initiaux pour les préparer à un vaste horizon d'usages possibles. » (Denis et Goëta, 2013, p. 16).

37. Traduction du terme *library technical workers* inspirée d'une dénomination utilisée au Canada, notamment par les bibliothèques montréalaises (@BiblioQC, 2020).

38. « We might call such contextual information *metadata,* or *metacontent,* or *metaknowledge*, but at any of these scales, the production and reproduction of metainformation for information storage, as well as the effective use of that metainformation at the point of retrieval, are both necessary for preserving not only the sense but also the value of the information from the old context to the new. » (Downey, 2014, p. 154).

le temps, à travers les disciplines ou à travers les cultures. Downey explique comment

> All this labor of moving data, content, or knowledge from one context to another [...] depends on understanding, manipulating, producing, and reproducing further descriptive, contextual information about that data, content, or knowledge. (Downey, 2014, p. 154)

Ainsi, bien que ces employés de bibliothèque n'aient pas souffert de la même invisibilité que les garçons messagers œuvrant dans le cadre du télégraphe, il s'agit de mettre en lumière « their historical and ongoing value over a long period of technological and economic transformation. » On retrouve ainsi cette notion de maintenance :

> If libraries are to remain relevant as tools of knowledge production and circulation, they must not only do their best to produce intelligible cataloging at the entry point of a collected item, but they must also continue to reproduce and repair that cataloging through the life of an item —- in fact, the life of an item in the library actually depends on the effectiveness of its cataloging, because no matter how theoretically valuable it may be, it is only actually valuable if it circulates. (Downey, 2014, pp. 153-154)

Plus récemment, toujours dans le contexte des bibliothèques, Lovins et Hillmann ont décidé de reprendre le cadre conceptuel issu des études en *maintenance, breakdown and repair* et de l'appliquer aux vocabulaires utilisés par les bibliothèques dans le contexte du Web sémantique. Relevant que ces vocabulaires prennent de plus en plus d'importance dans le contexte actuel, ils s'inspirent en particulier du travail de Jackson (2014) et de sa notion de *Broken World Thinking* :

> Where infrastructure is viewed not as a fundamentally sound system with occasional lapses, but rather as a fragile network of dependencies in various states of disrepair, kept from further collapse only by diligent attention and intervention of persons one might call « maintainers ». (Jackson, 2014, cité par Lovins et Hillmann, 2017)

Déclinant cette pensée à travers le concept *Broken Vocabulary Thinking*, Lovins et Hillmann souhaitent mettre en lumière le rôle des *maintainers* dans le contexte des vocabulaires[39] nécessaires au développement du Web sémantique. En effet, ils expliquent que si ce rôle est négligé, cela a pour

39. Lovins et Hillmann distinguent trois types de données bibliographiques : les *instance data* telles que les notices bibliographiques MARC ; les *structural vocabularies*, tel que RDA *Resource Description & Access* ; les *value vocabularies* tels que les vedettes-matières de la *Library of Congress* (LCSH).

conséquences des priorités peu équilibrées, où les investissements font la part belle aux projets innovants, au détriment des personnes et des protocoles assurant l'entretien et la maintenance, c'est-à-dire, l'innovation durable. Lovins et Hillmann espèrent donc qu'en participant à rendre ce travail de maintenance davantage visible [40], ils pourront attirer l'attention sur des tâches sous-financées, permettant ainsi aux vocabulaires de réaliser leur plein potentiel.

À la lecture de cet article, plusieurs niveaux de maintenance peuvent être discernés. Premièrement, Lovins et Hillmann expliquent que, d'une certaine manière, les vocabulaires bibliographiques ont toujours été *cassés*. En effet, le monde entourant ces vocabulaires descriptifs ne cesse de se *briser* et de changer (avec l'évolution de la compréhension de la condition humaine, les nouvelles découvertes scientifiques ou encore les changements géopolitiques tels que la dissolution de la Tchécoslovaquie en 1992). Si les bibliothécaires ne peuvent pas « réparer » ce monde extérieur, ils doivent toutefois maintenir les systèmes d'information permettant à leurs utilisateurs de mieux le comprendre (Mattern, 2018), quitte à se retrouver soumis à une injonction paradoxale :

> Library catalogers are expected to make rapid and hard-to-change decisions about how to organize texts in a way that is meant to serve future populations and needs that cannot reliably be known, balancing the intellectual depth and detail of their work (with greater-quality cataloging thought to bring greater long-term usability) with the very real economic and time cost of that work (with lower-quality cataloging thought to bring greater short-term savings). Lower-quality cataloging might save money in the short term, allowing the purchase of more books out of limited budgets, further pressuring libraries to spend less time (and money) on cataloging. (Downey, 2014, p. 154)

Deuxièmement, le travail d'harmonisation requis pour prendre en charge la multitude de standards et de vocabulaires utilisés dans le contexte des bibliothèques et du patrimoine culturel constitue un autre type de maintenance. Nous pouvons par exemple penser aux travaux de Zavalina et Zavalin, qui se sont penchés sur la façon dont les données d'autorité évoluent au cours du temps, et plus précisément en réaction à l'évolution des standards. Ils ont ainsi observé comment un corpus de 400 000 données d'autorité mises à disposition par l'OCLC a évolué dans un intervalle de 22 mois en fonction de

40. « But the narrative of innovation needs to change, so that activities mostly invisible today are made visible in the future » (Lovins et Hillmann, 2017).

l'évolution du nouveau standard de catalogage RDA[41] (Zavalina et Zavalin, 2018).

Enfin, Lovins et Hillmann expliquent que les bibliothécaires doivent faire évoluer leurs pratiques de *maintenance and repair* pour qu'elles s'adaptent à l'implémentation de leurs vocabulaires sur le Web. Cet ajustement nécessite de mettre à jour les flux de travail préexistants : « in order to be sustainable on the Web, these workflows need updating to be as distributed and machine-actionable as possible. » Envisager la publication de ces vocabulaires sur le Web à travers le prisme du *Broken World Thinking* signifie donc établir de nouvelles priorités :

> Recognition of the fragility of current systems and preparation for inevitable breakdowns ; building maintenance functions directly into tools, workflows, and budgets, and including documentation, preservation, and terms of use from the moment projects are conceived. (Lovins et Hillmann, 2017)

Concrètement, cela se traduit par l'adoption de nouveaux outils tels que des logiciels de gestion de versions décentralisés tel que Git proposé par la plateforme en ligne GitHub, mais également par un changement de posture :

> One of the challenges in managing this kind of distributed content is moving from a traditional *filter, then publish* to *publish, then filter*. The publish-then-filter model allows any authorized person to add new information, after which it will be discussed, edited, enhanced. (Lovins et Hillmann, 2017)

À la même période, Arnold (2016) s'est intéressé à la notion de maintenance dans le contexte des archives, affirmant que les archivistes sont tous des *maintainers* : « we do the hard and invisible work of maintaining records. [...] We maintain memory and accountability ». Convaincu que la *maintenance theory* peut aider le milieu à fixer le problème d'invisibilité dont il souffre, il en a fait le sujet de son discours lors d'une rencontre de la Society of American Archivists. Énumérant les différentes formes que prend cette invisibilité (profession peu considérée et sous-payée[42], contrainte de faire appel à des stagiaires et des intérimaires pour faire face à la pression de faire plus avec moins ; travail parfois *effacé* par les archivistes eux-mêmes, par souci d'impartialité et de professionnalisme ; travail émotionnel indis-

41. Et relèvent notamment que : « the change in application of certain data elements, related to evolution of RDA standard, was observed, with gradual and sometimes drastic increase in the use of elements representing persons, as well as some of the Linked Data-enabling elements. Despite the observed growth, the level of application of Linked-Data enabling elements in authority records remains relatively low » (Zavalina et Zavalin, 2018, p. 596).

42. « Our work is misunderstood, undervalued, often taken for granted. » (Arnold, 2016).

pensable pour maintenir de bons réseaux de relations [43] mais passant inaperçu ; tâches de maintenance (gestion des collections, préservation, transports, etc.) éclipsées au profit de l'innovation et des *digital archivists* et bien souvent réalisées dans le cadre de contrats à temps partiel ou à durée déterminée), il encourage la profession à prendre de la distance face à la distribution du pouvoir et de la visiblité favorisant *innovators and disruptors* et à se demander : *who does maintenance work, when and where and why* [44] et, partant de cela, à se battre pour ses droits.

Si Arnold aborde la question de la maintenance d'un point de vue *macro*, il est également possible d'examiner cette question à un niveau plus *micro*. En effet, la conservation, la préservation et l'ouverture des archives à la recherche impliquent bien sûr également des tâches de maintenance au niveau des données elles-mêmes.

Ainsi, au cours des années 1980 déjà, Bearman remarque que le contrôle d'autorité commence à faire l'objet d'une attention accrue au sein de la communauté archivistique. Cette dernière s'y intéresse depuis quelques années, après avoir été sensibilisée aux problèmes de conception des systèmes et de qualité des données grâce à ses expériences dans les applications d'automatisation et à sa participation à des bases de données bibliographiques nationales (Bearman, 1989). Bearman rend toutefois attentif au coût qu'entraîne la maintenance d'un vocabulaire contrôlé : « [Authority control involves] substantial intellectual, technical, and administrative overhead. » (Bearman, 1989, p. 287). Il étaye son propos en citant des recherches ayant montré qu'en raison de toutes ces exigences, « many systems that supposedly employ authority control, especially those dependent on manual updating, break down and thereby fail to deliver on their promise » (Palmer, 1986 et Thomas, 1984, cités par Bearman).

Plus récemment, dans le contexte du projet *schema.org* – qui œuvre à la structuration de données minimales destinées à décrire les sites web – et des ajouts proposés par le *W3C Schema Architypes Community Group* [45], c'est toujours cette même maintenance qui est évoquée pour justifier un certain choix de modélisation, même si les technologies ont évolué :

> In theory we could have introduced archive types for each type thing you might find in an archive, such as ArchivedBook, ArchivedPhotograph, etc. – obvious at first but soon gets difficult to

43. Avec les producteurs d'archives, les personnes « sujets » d'archives, les utilisateurs finaux, les donateurs et des communautés plus larges telles que la communauté archivistique (Arnold, 2016).

44. Questions issues de Downey (2014).

45. La mission de ce groupe est de discuter et de préparer des propositions d'extension du schéma Schema.org afin de favoriser une meilleure représentation des archives numériques et physiques et de leur contenu, voir : https://www.w3.org/community/architypes/.

> scope and maintain. Instead we took the MTE [Multi Typed Entity] approach of creating a type (ArchiveComponent) could be added to the description of any thing, to provide the archive-ness needed. (Wallis, 2019)

Arnold explique que le travail de maintenance est souvent invisible, du moins jusqu'à ce que quelque chose se casse : « When it does, the maintainers have to clean up the mess, whether it's a tree that fell on power lines, garbage that wasn't picked up, or a levee that was breached » (Arnold, 2016).

Or, les *maintainers* ne sont pas toujours en mesure d'agir en profondeur sur les dysfonctionnements. Nous pouvons par exemple penser à la maintenance des logiciels informatiques, derrière lesquels se cache une accumulation de couches de code, comme autant de pansements et de béquilles dégainés pour faire face à la survenue de bugs ou à la demande d'ajouts de nouvelles fonctionnalités dans un délai serré. Dans de telles configurations, les gains à court terme sont alors favorisés au détriment de développements pensés sur le long terme [46], et génèrent ce que Cunningham a illustré en 1992 à l'aide de la métaphore de la *technical debt*. Cette dette est accompagnée d'intérêts, à savoir les efforts supplémentaires de maintenance qui découlent de cette dette technique [47]. Or, si elle n'est pas prise en charge, ces derniers ne vont cesser d'augmenter (Cunningham, 1992).

Alors que la métaphore initiale de Cunningham se concentrait sur les problèmes de qualité de code, la littérature à ce sujet s'est depuis développée et, en 2015, Li *et al.* ont relevé pas moins de dix types de dette technique [48] au cours de leur analyse de près d'une centaine de publications à ce sujet. Des outils et métriques ont par ailleurs été développés, à l'instar du *repair effort*, destiné à quantifier la dette technique et, par conséquent, le coût nécessaire pour améliorer la qualité d'un logiciel (Nugroho *et al.*, 2011). En combinant ce coût à une estimation des efforts de maintenance dus à une piètre qualité technique, il devient ainsi possible de calculer le retour sur investissement d'un *repair work* et d'adapter le niveau de qualité visé en fonction de ces estimations. Cela permet également de déterminer quelle stratégie sera la plus avantageuse entre l'amélioration de code préexistant ou sa réécriture complète, en tenant également compte de facteurs tels que l'expérience de l'équipe (Nugroho *et al.*, 2011). De tels calculs peuvent également participer

46. Par exemple en prenant le temps de rédiger une documentation adéquate ou en procédant à un *réusinage* du code.

47. Voir l'illustration de cette métaphore proposée par Martin Fowler pour une meilleure visualisation du problème : https://martinfowler.com/bliki/TechnicalDebt.html/

48. Requirements debt ; architectural debt ; design debt ; code debt ; test debt ; build debt ; documentation debt ; infrastructure debt ; versioning debt ; defect debt (Li *et al.*, 2015).

à cibler certaines zones d'attention susceptibles de faire l'objet de mesures préventives à l'avenir.

Dans le cadre de ses recherches sur la maintenance des bases de données, Cleve a ainsi mis en exergue l'impact que peut générer un manque de documentation. Il estime ainsi que l'étape de compréhension d'un système logiciel, connu sous le nom de *reverse engineering*, peut constituer jusqu'à 80% du total des efforts de maintenance (Cleve, 2016). Si ces préoccupations concernant le principe de dette technique pourraient paraître accessoires au premier regard, elles apparaissent toutefois nécessaires et stratégiques dès lors qu'on prend connaissance des coûts colossaux que représente la maintenance logicielle. Ainsi, l'historien Nathan Ensmenger estime qu'entre le début des années 1960 à nos jours, ces coûts ont représenté entre 50% à 70% de toutes les dépenses de développement logiciel (Ensmenger, 2016).

Clair propose d'appliquer cette métaphore de la dette technique au contexte des bibliothèques. Constatant que les archives et les bibliothèques doivent faire face à la gestion de plus en plus de contenu numérique, résultant en d'important coûts de maintenance des métadonnées décrivant ce contenu [49], il avance que ces métadonnées pourraient être considérées comme le *codebase* nécessaire au bon fonctionnement des services techniques des bibliothèques et que « the labor, or lack thereof, required to ensure sufficient metadata for a properly functioning system can be thought of as a down payment toward the relief of that technical debt » (Clair, 2016).

S'inspirant des typologies de Tom *et al.* (2013) et de Li *et al.* (2015), il distingue cinq types de dettes techniques pouvant être identifiées dans le contexte de la gestion des métadonnées d'une bibliothèque et pouvant survenir tant de façon intentionnelle que non intentionnelle : « code debt ; design and architectural debt ; environmental debt ; documentation debt ; requirements debt » [50].

Si la contribution de Clair est avant tout théorique et composée de recommandations demeurant très générales [51], il relève l'opportunité de pouvoir étudier la dette technique à travers des études de cas dans le cadre de la conversion des métadonnées des bibliothèques vers de nouveaux standards de catalogage comme BIBFRAME et RDF.

C'est dans ce contexte de RDF que Magnus Salgö, Wikimédien suédois travaillant sur les relations entre GLAMs et Wikidata, utilise également le

49. « In the form of cleanup of purchased vendor records, conversions of metadata to different representations, and migrations across systems and applications. » (Clair, 2016).

50. Cette typologie composée de cinq types de dettes sera détaillée et réutilisée dans le cadre de l'étude de cas présentée en seconde partie de cette thèse.

51. Il conseille par exemple aux bibliothèques désireuses de limiter leurs dettes techniques de considérer « an ongoing program of metadata reviews for quality, ability to meet user needs, and suitability for various maintenance tasks (migrations, conversions to new data models and/or standards, etc.) » (Clair, 2016).

terme de *metadata debt*[52] ; plus précisément pour désigner les problèmes de qualité dus à la présence de métadonnées descriptives disponibles seulement sous forme de texte littéral et par conséquent soumises à des problèmes d'ambiguïté :

> I see we have bad curated archives delivering RDF but use « strings[,] not things » and we have to pay a prize of not finding what we are looking for [...] (@Salgo60, 2019).

Nous envisagons donc cette forme de dette technique comme un sous-ensemble de la catégorie *code debt* décrite par Clair. En effet, à l'instar du code source d'une application qui ne serait pas écrit de façon optimale et accroîtrait les efforts de maintenance requis pour assurer le bon fonctionnement de cette application, les métadonnées descriptives – telles que des noms de lieux, par exemple – contenant du texte littéral non compréhensible par une machine vont augmenter les efforts de maintenance nécessaires à l'avenir[53]. Le prix à payer pour l'utilisateur final sera une recherche non optimale pouvant générer du bruit ou du silence[54], tandis que les intérêts à payer par l'institution se manifesteront par un temps de traitement supérieur de ces métadonnées. Par exemple, toute velléité de géolocalisation des données à travers une application dédiée nécessitera au préalable un travail de nettoyage et de désambiguïsation, afin de pouvoir associer à des chaînes de caractères un identifiant unique et persistant auquel pourront être associées des coordonnées géographiques.

52. Dans un article de blog dédié, il le définit comme « the cost of getting your data usable because of lack of good metadata » (Sälgö, 2020).

53. Soulignons toutefois que la notion de qualité des métadonnées s'étend bien au-delà de ce caractère structuré et lisible par des machines. Comme Bruce et Hillmann ont par exemple eu à cœur de le décrire voilà plus de 15 ans déjà – peu après l'apparition du protocole d'échanges de métadonnées OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) dont l'implémentation participa à révéler les difficultés propres à l'agrégation de données hétérogènes –, plusieurs dimensions (*completeness ; provenance ; accuracy ; conformance to expectations ; logical consistency and coherence ; timeliness ; accessibility*) permettent en effet de définir et dès lors de mesurer la qualité des métadonnées. Le caractère structuré et lisible par des machines, spécifique au contexte du Web de données dans lequel s'inscrit cette thèse, ne représente donc qu'une facette de cette qualité et renvoie aux dimensions plus globales de l'adéquation aux usages et de l'accessibilité.

54. Ce qui nous ramène au constat de Downey qui observait que lorsque les bibliothécaires confrontés à des coupes budgétaires optent pour un catalogage de moindre qualité, cela permet certes de plus grandes économies à court terme, mais menace la qualité de l'accès aux collections sur le long terme et condamne les bibliothèques à consacrer de moins en moins de temps et d'argent au catalogage (Downey, 2014, p. 154).

Ce type de dette correspond à ce que Clair décrit comme un « excess of sub-standard metadata requiring maintenance on the part of a cataloger », le standard étant ici celui du Web sémantique, à savoir des données structurées lisibles par des machines désignées à l'aide d'*URIs (Uniform Resource Identifiers)*, c'est-à-dire de courte chaîne de caractères identifiant une ressource physique ou abstraite sur un réseau (Berners-Lee *et al.*, 2005). Nous choisissons d'introduire le terme de *dette sémantique* pour désigner dans les pages qui suivent ce type de dette.

S'il est urgent de prendre des mesures dans le cadre de l'encodage de nouvelles données pour stopper l'accroissement de cette dette technique, il n'en reste pas moins que le traitement de cette dette elle-même nécessite que des ressources y soient consacrées. Or, comme on l'a vu dans le cadre de la *maintenance theory*, ce travail se caractérise par son invisibilité et est dès lors plus difficile à valoriser que des projets surfant sur la vague de l'innovation.

Le 7 août 2019, Kelly David, travaillant sur la standardisation du *Getty Provenance Index*, haranguait ainsi les utilisateurs de Twitter :

> how do you justify the need for funding research and unsexy metadata work ? I get it that it doesn't appeal to the higher ups. but it. is. NECESSARY. work. you want sexy platform ? you need lots of unsexy data cleaning (@daviskellyk, 2019)

Ce à quoi Stuart Myles, *Director of Information Management at Amazon Web Services*, réplique en expliquant qu'en adaptant le concept de *technical debt* aux métadonnées [55], il devient possible de mettre l'emphase sur le temps et l'argent en jeu :

> we did'nt make a « mistake » before, we made a trade-off (we took on debt to go faster). and we can put off paying the debt if we want, but there is a cost (we keep paining more-n-more interest). (@smyles, 2019b)

Bien qu'un discours mettant en lumière les coûts occasionnés par ce type de dette soit essentiel et qu'une meilleure valorisation des activités de maintenance des métadonnées soit hautement souhaitable, nous souhaitons pro-

55. « [I] have tried to adapt the concept of "technical debt" [...] into "metadata debt" : we've cut some corners with our metadata, in order to get things done quick-n-dirty. but now we've built up metadata debt, which means we need to pay interest on it, or pay it off » (@smyles, 2019a).

fiter de cette thèse pour explorer de façon empirique les opportunités dont disposent des institutions dans leur travail quotidien de maintenance. Pour ce faire, nous nous référons au principe de *good enough*, sous-jacent à la thèse exposée par Greene et Meissner dans cet article publié en 2005 qui a eu beaucoup de retentissements au sein de la communauté archivistique : *More Product, Less Process (MPLP)*.

Dans un contexte de ressources limitées, d'accroissement du volume des fonds à traiter et d'attentes grandissantes des utilisateurs, ils exhortent les services d'archives à favoriser un accroissement de l'accès aux fonds par le biais de *minimal processing*, plutôt qu'un accès optimal à une minorité seulement de collections. En effet, face au constat sans appel que « Cataloguing is a function which is not working » (Best Value, 2003, cité par Greene et Meissner, 2005), ils voient deux options possibles :

> One is to increase the resources devoted to it. Given all we know about current processing practices, current acquisition levels, and current backlogs, it would require roughly a tripling of the number of processing archivists to fix the problem in this way. Is there anyone willing to suggest with a straight face that this is possible ? The other option is to change the way we process so that we can, with our existing resources, roughly triple the speed with which we process. (Greene et Meissner, 2005, p. 254)

Ils proposent dès lors une série de nouvelles lignes directrices pour le classement, la conservation et la description des fonds d'archives [56] :

1. Expediting the availability of collections to users ;
2. Assuring adequate arrangement of materials for users' needs ;
3. Taking the minimum steps necessary for physically preserving collection materials ;
4. Describing materials sufficiently for use.

Si les auteurs prônent une telle approche – que nous pouvons par ailleurs rapprocher du concept de *fitness for use* (Boydens et van Hooland, 2011) sur lequel nous reviendrons au cours du chapitre 3 – c'est avant tout pour pouvoir faire face à l'important volume de fonds n'ayant pas encore été pris en charge *(backlogs)* auquel doivent faire face les institutions [57]. Comme ils l'expliquent :

56. L'article utilise les notions d'*arrangement, preservation, description*.

57. « Put very simply, processing is not keeping up with acquisitions and has not been for decades, resulting in massive backlogs of inaccessible collections at repositories across the country (and across all types of archival institutions). It should be dismaying to realize that our profession has been struggling with backlogs for at least sixty years. [...] These backlogs are continuing to grow. And they are weakening the archival profession. » (Greene et Meissner, 2005, pp. 208-209).

> MPLP is simply the archival incarnation of the Principle of Good Enough, a software engineering rule that promotes quick, simple, and extensible designs over more elaborate and costly ones. It is very much in this spirit that we believe MPLP to have enduring value for archives and for related professional communities. (Meissner et Greene, 2010, p. 218)

La présente thèse cherche donc à s'inspirer de cette posture, plus qu'elle n'aspire à suivre strictement les préceptes énoncés par Greene et Meissner au sujet de la description et de l'indexation des archives. En effet, bien que ces derniers témoignent d'une lucidité salutaire [58], il ne s'agit pas tant ici d'agir sur la description des fonds d'archives n'ayant pas encore pu être inventoriés et indexés, ou de redéfinir les approches actuelles d'indexation, que de reprendre la stratégie de *minimal processing* déployée pour faire face aux *backlogs* et de réfléchir à ce qu'une telle approche impliquerait dans le contexte de la maintenance des métadonnées archivistiques et plus précisément des données d'autorité.

En effet, comme le souligne Arnold, la méthodologie MPLP peut être considérée comme :

> A framework that can (and should) be used as a tool to empower archival maintainers who process collections by encouraging them to employ professional judgement toward what was previously considered rote and formulaic work. (Arnold, 2016)

Il semble dès lors opportun de réfléchir à la manière dont cette approche de *good enough* pourrait venir éclairer la question de la maintenance des données d'autorité archivistiques, et par extension, celle du traitement de la dette générée par des problèmes de qualité antérieurs.

Tendre vers un idéal de *good enough* plutôt que vers une perfection ne pouvant être atteinte que pour une minorité de données – au détriment du plus grand nombre –, pourrait par exemple consister à mettre en place des processus automatisés permettant d'aligner des noms de personnes (ce qui implique de devoir tolérer un certain degré d'incertitude quant au fait qu'il s'agisse véritablement d'un même individu), ou encore se reposer sur des données mises en ligne par d'autres organismes (sans pour autant être en

58. Comme en témoigne par exemple cet extrait : « Although description is obviously a critical archival function that serves multiple roles, it needn't be long-winded, laborious, or minutely detailed to be effective. A crisp, simple presentation with minimal verbiage often provides the most effective representation of collection materials. [...] An unfortunate tendency on the part of processing archivists is to use the preparation of these text notes as an excuse to demonstrate their own knowledge (of both collection and historical context) and writing ability. Perhaps this is an attempt to demonstrate professionalism but, if so, it is a misguided one that further reduces processing productivity. » (Greene et Meissner, 2005, pp. 246-247).

mesure de vérifier systématiquement leur exactitude). Mais attention, le fait de chausser ces lunettes offrant une vision centrée sur le *good enough* ne signifie pas pour autant brader toutes les exigences et standards en vigueur. Comme l'ont souligné Lovins et Hillmann, c'est en reconnaissant l'importance des normes, des protocoles et, dès lors, de la maintenance, qu'il est possible de prétendre à une innovation soutenable [59] :

> We can go beyond (though not leave behind) the never-ending cycle of breakdown, maintenance, and repair, but it requires that we take very seriously the need to invest in the community of maintainers and tools that keep bibliographic vocabularies available and sustainable over time. (Lovins et Hillmann, 2017)

Cette thèse apparaît donc comme une opportunité d'examiner sur quels aspects il est possible d'agir pour mieux tirer parti de l'existant, tout en favorisant une gestion soutenable des données d'autorité. C'est ce que nous précisons dans les paragraphes suivants qui présentent nos questions de recherche.

59. *Sustainable innovation* dans le texte.

Questions de recherche

La principale question de recherche de cette thèse est la suivante :

> *Comment favoriser une gestion soutenable des données d'autorité archivistiques dans le cadre du Web de données ?*

Avant de poursuivre avec l'introduction des sous-questions de recherche permettant d'explorer cette question, commençons par décortiquer les composantes de cette question :

Comment favoriser Il s'agit d'examiner à l'aide de données empiriques quelles méthodes et quels outils pourraient faciliter, conforter, seconder, soutenir...

Une gestion soutenable Nous choisissons d'utiliser l'expression *gestion soutenable*, bien que l'adjectif le plus usité dans le monde francophone soit *durable* et non pas *soutenable* [60]. En effet, des voix se sont élevées au sein des milieux traitant de l'environnement et de la gestion des ressources naturelles disponibles, pour promouvoir l'usage du terme *soutenable*, une traduction littérale du terme anglophone qui serait plus à même d'évoquer les changements devant être mis en place pour tenir compte de la finitude des ressources [61]. Même si les ressources prises en compte dans le cadre de cette thèse sont avant tout humaines, financières et organisationnelles, cette prise de position garde toute sa pertinence : il s'agit de favoriser une gestion pouvant être soutenue sur le long terme, en tenant compte des ressources disponibles. Par ailleurs, bien qu'il ne s'agisse pas d'un axe de recherche que nous aurons l'occasion d'approfondir dans le cadre de cette thèse, il est clair que la question des enjeux environnementaux ne peut être ignorée. En effet, la croissance exponentielle des données numériques s'accompagne de conséquences environnementales qui requièrent une prise de recul sur la façon dont les données sont stockées et interrogées, ainsi que sur les éventuelles mesures pouvant être prises pour limiter ces conséquences.

60. Comme l'illustre sans équivoque cette représentation Ngram Viewer (application qui montre l'évolution, dans des sources imprimées, de la fréquence d'un ou plusieurs mots à travers le temps) : https://tinyurl.com/y45zqaf2.

61. L'ethonologue Thierry Sallantin l'expose sans langue de bois : « dire *durable*, c'est faire injure à la langue anglaise qui possède le mot *durable* dans sa langue et ne l'a pas choisi, préférant *sustainable*, qui a une longue tradition d'usage dans le vocabulaire anglais pour traiter des sciences de la gestion des forêts. [...] C'est faire aussi injure à l'origine française, attestée [...] de 1346, du mot *sustainable* (soutenable, soutenir, soutenabilité) » (Sallantin, 2012). (À noter que d'après le dictionnaire du Centre National de Ressources Textuelles et Lexicales, l'origine du terme, issu de *soustenable (que l'on peut supporter, endurer)* remontrerait au XIII^{e} siècle déjà (1246) (Centre national de ressources textuelles et lexicales, 2020).).

Des données d'autorité archivistiques Comme les prochaines pages le montreront en détail, cette thèse porte plus particulièrement sur les données d'autorité utilisées dans un contexte archivistique.

Dans le cadre du Web de données Cette thèse, rédigée en l'an 2020, est le fruit d'un contexte spécifique : elle tient compte des récents développements du Web et notamment de l'essor de bases de connaissance reposant sur des données structurées pouvant lier des ressources contenues jusque-là dans des silos.

Trois sous-questions de recherche sont déployées pour explorer cette piste. La première porte sur la qualité des données d'autorité préexistantes et plus précisément sur les possibilités de diminution de la *dette sémantique* dont elles font l'objet, pour éviter que les *intérêts* continuent à s'accumuler.

> *QR1 : Dans quelle mesure est-il possible de se reposer sur des processus d'automatisation pour réduire la dette sémantique pesant sur les données d'autorité ?*

Il s'agit de commencer par prendre en compte la dette technique plutôt que la laisser s'accumuler, de même que les *intérêts* qu'elle engendre. En effet, comme l'a souligné Clair (2016), le passage vers RDF pour la description des collections représente une opportunité pour effacer la dette technique existante. Il précise toutefois qu'il faut rester vigilant et que cela peut également être une source de dette technique supplémentaire (par exemple si les flux de travail ne sont pas bien documentés ou si les nouveaux modèles et normes ne sont pas connus de façon uniforme au sein de l'institution). Dans le contexte des données d'autorité destinées à être publiées et partagées sur le Web, cela consiste à travailler sur la qualité des données (structuration, standardisation, dédoublonnage), en accordant une attention particulière à leur sémantisation, ce processus visant à lever l'ambiguïté sur le sens d'un mot [62].

Aborder cette question de recherche à l'aide d'une approche *good enough* (voir paragraphes précédents) se traduit ici par le recours à des logiciels et algorithmes permettant d'automatiser une partie des processus, quitte à devoir composer avec un certain degré d'incertitude. La seconde question de recherche s'intéresse quant à elle aux possibilités offertes par Wikibase, le

62. En linguistique, le terme désigne l'« interprétation des éléments d'un texte » (Wiktionnaire, 2020). Il n'est pas aisé de trouver une définition en français de ce terme dans le contexte du Web sémantique. Dans le cadre du projet Histoire des arts' Lab, le Ministère français de la Culture parle de sémantisation de mots-clés par le biais de la production de *tags sémantiques*. Le *tagging sémantique* est décrit ainsi : « une indexation par concept issu d'un référentiel. Il s'agit d'expliciter le sens des informations afin que les machines puissent les exploiter de façon automatique, sans ambiguïté et à grande échelle », afin de pouvoir « lever l'ambiguïté sur le sens d'un mot » (Ministère de la Culture, 2016).

logiciel libre derrière Wikidata permettant de déployer de sa propre base de connaissance.

> *QR2 : De quelle manière les fonctionnalités offertes par le logiciel Wikibase peuvent-elles être utilisées pour faciliter et rationaliser le travail de création et de maintenance des données d'autorité ?*

Cette seconde question vise à explorer comment il est possible de soutenir et optimiser les activités de maintenance en déployant des outils et des mesures adaptées au contexte du Web sémantique. Comme le suggèrent Lovins et Hillmann, d'anciens processus de maintenance de données d'autorité ou vocabulaires contrôlés doivent être mis à jour : « in order to be sustainable on the Web, these workflows [63] need updating to be as distributed and machine-actionable as possible. » Ils préconisent ainsi d'inclure, dès la phase de conception d'un projet, ces fonctions au sein des outils, des processus de travail et des budgets, en s'inspirant d'exemples comme GitHub ou Wikipedia : « [they] illustrate what it looks like to have maintenance tasks *baked-into* regular operations. Much of this happens without conscious effort, thanks to automated versioning and roving *bots* » (Lovins et Hillmann, 2017).

Dans le cadre de cette thèse, il s'agit d'examiner les fonctionnalités et les extensions du logiciel Wikibase et de voir comment elles peuvent satisfaire les besoins d'une institution en ce qui concerne la gestion des données d'autorité, mais également d'aller plus loin et d'observer, toujours dans une optique de *good enough*, ce qui pourrait éventuellement permettre de rationaliser certaines étapes ou de faire appel à l'intelligence collective, par exemple par le biais de l'édition collaborative [64]. Enfin, en ce qui concerne la troisième question de recherche, elle vise à examiner comment il est possible de mutualiser la production de données d'autorité, en particulier dans le cadre du Web de données.

> *QR3 : Comment les Linked Open Data peuvent-elles faciliter de nouvelles formes de mutualisation susceptibles de réduire le volume de données à maintenir ?*

En effet, comme Bearman le suggérait en 1989 déjà,

> Not all reference files used in archival information systems need to be created *de novo* by archivists. Reference files in an archival

63. « Including centralized, human-centered quality assurance, publishing, and versioning ».

64. Comme l'expliquent Lovins et Hillmann, « the publish-then-filter model avoids the bottleneck of editorial approval, allowing any visitor to find new content, spot errors or omissions, and correct them in real time (if authorized), and become new members of the community » (Lovins et Hillmann, 2017).

> information system may be the primary databases of the discipline or organization that created them. Archivists and the builders of other cultural information systems only need identify databases that contain information which could be linked to records, and then import such databases into their systems. [...] Because reference files can be acquired from other disciplines, especially from the computerized databases of the parent organization of the archives, employing a deep network of reference files for one application need not be prohibitively costly. It does, of course, require implementations that can support multiple, independent, and conflicting authorities. (Bearman, 1989, p. 297)

Cette troisième question est donc l'occasion d'étudier à l'aide de cas concrets de quelle manière les données ouvertes et liées peuvent faciliter le partage et la réutilisation de *fichiers de référence* en 2020. Outre les possibilités d'alignement et d'enrichissement, il s'agit de tester – à l'aide de Wikibase et des requêtes SPARQL fédérées –, les possibilités de gestion semi-centralisée combinant des données gérées localement et des données maintenues par d'autres institutions. Cela permettrait ainsi aux archivistes occupés à décrire des fonds de se concentrer sur la maintenance de données de qualité concernant leur cœur de métier, à l'instar des bibliothécaires qui ont pris l'habitude de mutualiser leurs données de référence (Johannic-Seta, 2017; Pouchol, 2016).

Le tableau 1 propose une vue récapitulative des trois questions de recherche, accompagnées du principal enjeu associé à chacune d'entre elles et de la contribution attendue dans le cadre de cette thèse.

Ce sont ces trois questions de recherche qui constituent la charpente principale de cette thèse : elles sont d'abord abordées de façon transversale dans le cadre de la première partie dédiée à l'état de l'art, avant d'être traitées de manière empirique à l'aide d'une étude de cas.

La première partie est construite comme suit : le premier chapitre porte sur l'évolution et la stucturation progressive des données d'autorité archivistiques relatives à des personnes physiques, des index onomastiques jusqu'aux URIs propres au Web sémantique, en s'intéressant notamment aux initiatives de mutualisation des données d'autorité, tandis que le second chapitre analyse l'usage qui est fait du logiciel Wikibase, en particulier dans le secteur du patrimoine culturel.

La seconde partie débute par un chapitre présentant l'étude de cas et de son contexte ; viennent ensuite les chapitres dédiés au traitement et à la sémantisation des données d'autorité, puis à la mise en œuvre d'une Wikibase, de la création de propriétés aux possibilités de mutualisation des données ; ils sont suivis d'un chapitre proposant une analyse SWOT et des

Question	**Enjeu**	**Contribution attendue**
QR1	Réduction de la *dette sémantique*	Possibilités et limites des processus semi-automatisés pour la structuration, sémantisation et réconciliation des données nominatives
QR2	Rationalisation de la maintenance des données d'autorité	Possibilités et limites du logiciel Wikibase
QR3	Mutualisation de la maintenance des données d'autorité	Possibilités et limites du partage de données dans le cadre du *LoD* cloud

TABLE 1 – Vue récapitulative des trois questions de recherche, enjeux et contributions attendues.

recommandations généralisables. Enfin, la thèse se clôture avec la formulation de conclusions et de perspectives pour l'avenir.

Méthode

Comme évoqué en fin de section précédente, cette thèse en Sciences et technologies de l'information et de la communication est construite autour de deux parties distinctes : la première partie est dédiée à un état de l'art critique, tandis que la seconde s'appuie sur une étude de cas. Cette section détaille nos choix de méthode et les caractéristiques de cette recherche [65].

La méthode adoptée dans le cadre de cette thèse combine une approche *bottom-up* et une approche *top-down*. En ce qui concerne l'approche *bottom-up*, nous nous référons aux travaux du linguiste Blanchet et à ce qu'il qualifie de *méthode empirico-inductive* [66], c'est-à-dire une méthode permettant au chercheur de comprendre un phénomène à partir d'un ensemble de données plutôt que d'évaluer un modèle ou une hypothèse préétablie à l'aide de données collectées à cette fin (Blanchet, 2000, p. 31). Une telle méthode se caractérise par le fait que :

> D'une certaine façon, les données priment sur la construction intellectuelle, tant en termes de déroulement du travail que, surtout, de méthode d'enquête et de traitement de ces données, puisque l'interprétation produite est toujours relative aux données, dont elle émerge. (Blanchet, 2000, pp. 31-32)

Si, comme le rappelle le linguiste, « il s'agit de *comprendre* (c'est-à-dire de "donner du sens à des événements spécifiques") et non d'*expliquer* (c'est-à-dire d'établir des lois universelles de causalité) » (Blanchet, 2000, pp. 30-31), il faut toutefois savoir que cette méthode peut être combinée avec une approche de type *top-down*, dans un « va-et-vient inductif/déductif » (Blan-

65. Les détails plus opérationnels, tels que les manipulations et traitements de données sont présentés dans le corps du document, dans des sections dédiées du chapitre 4.

66. Comme le précise Blanchet, « on pourrait penser qu'une démarche est *déductive* lorsqu'elle s'appuie avant tout sur l'observation des phénomènes et *inductive* lorsque l'hypothèse induit un expérimentation. En fait, c'est l'inverse : les phénomènes empiriques sont considérés comme induisant une théorie lorsqu'ils la précèdent et une expérimentation *ad hoc* se *déduit* d'une théorie *a priori*. » (Blanchet, 2000, p. 34).

chet, 2000, p. 32), notamment dans le cadre de l'étude de cas [67]. Notre recherche s'appuie sur une telle configuration.

Le choix de notre cas [68] ayant été motivé par les premiers constats réalisés dans le cadre de notre participation au projet de recherche Adochs [69], c'est l'approche de type empirico-inductive qui a été utilisée pour formuler une hypothèse de recherche à partir de nos observations issues du terrain, plutôt que pour vérifier des hypothèses préalables. En effet, comme l'a démontré Flyvbjerg, « the case study is useful for both generating and testing of hypotheses [...] » (Flyvbjerg, 2006, p. 425).

Précisons toutefois que si le choix de ce cas a été inspiré par une opportunité, il n'en est pas moins réfléchi et argumenté : comme mis en évidence par Gagnon, l'étude de cas unique est en effet recommandée « pour une problématique de type empirique brut, c'est-à-dire un phénomène jusque là inexploré » (Gagnon, 2005, p. 43). Or, dans notre situation, les premiers contacts avec le terrain du cas pressenti furent à la fois le révélateur d'un phénomène complexe peu exploré jusque-là [70] qu'il semblait pertinent d'étudier [71], et une confirmation du fait que l'institution et ses particularités s'y prêtaient. En effet, comme l'a montré Flyvbjerg – qui s'est attaqué aux conceptions simplistes des études de cas véhiculées par la sagesse populaire [72] –, le caractère généralisable d'une étude de cas peut être accru par la sélection stratégique du ou des cas. Il a en effet remarqué que :

> When the objective is to achieve the greatest possible amount of information on a given problem or phenomenon, a representative case or a random sample may not be the most appropriate strategy. This is because the typical or average case is often not the richest in information. Atypical or extreme cases often reveal more information because they activate more actors and more basic mechanisms in the situation studied. (Flyvbjerg, 2006, p. 425)

67. Elle est définie comme « une enquête empirique qui étudie un phénomène contemporain dans son contexte de vie réelle, où les limites entre le phénomène et le contexte ne sont pas évidentes, et dans laquelle des sources d'information multiples sont utilisées. » (Yin, 1984, p. 23, cité par Blanchet, 2012).

68. En l'occurence, le Centre d'Étude et de Documentation Guerre et Sociétés Contemporaines (CegeSoma), une présentation détaillée en est faite en seconde partie de thèse.

69. Ce projet BRAIN, financé par la Politique scientifique fédérale belge a débuté en novembre 2016. Il vise à améliorer les processus de contrôle de qualité des collections patrimoniales numérisées, voir : http://adochs.be.

70. La gestion des données d'autorité archivistiques dans le cadre du Web de données.

71. Comme le concluent Harrison *et al.* dans leurs travaux de synthèse sur les études de cas, « Case study research can be used to study a range of topics and purposes however, the essential requisite for employing case study stems from one's motivation to illuminate understanding of complex phenomena » (Harrison *et al.*, 2017).

72. Comme par exemple le fait que « one cannot generalize on the basis of a single case or that case studies are arbitrary and subjective » (Flyvbjerg, 2006, p. 432).

Ainsi, parmi les différentes stratégies que le chercheur énumère pour sélectionner un cas, notre attention s'est portée sur le cas critique (*critical case*), dont l'objectif est résumé ainsi :

> To achieve information that permits logical deductions of the type « if this is (not) valid for this case, then it applies to all (no) cases ». (Flyvbjerg, 2006, p. 426)

Le cas pressenti pour cette recherche – le CegeSoma – relève de cette catégorie dans la mesure où, comme nous le verrons, il s'agit d'une institution de taille modeste aux ressources limitées, jouissant d'une infrastructure informatique vieillissante [73]. Nous pouvons dès lors considérer que ce qui fonctionne pour un centre de recherche et de documentation de taille modeste pourra être généralisé à des cas similaires, mais également à des cas bénéficiant d'infrastructures et de moyens plus conséquents [74].

En ce qui concerne la démarche adoptée dans le cadre de cette étude de cas, elle découle des conditions de réalisation de notre doctorat qui nous amènent, par le biais du programme BRAIN [75], à intégrer à temps partiel [76] un établissement scientifique fédéral [77], durant une période de quatre ans. De telles collaborations, également à l'œuvre dans le cadre français des dispositifs de Conventions Industrielles de Formation par la Recherche (CIFRE), représentent une opportunité de contribuer à la recherche scientifique « dans un contexte d'application pratique, mais également transdisciplinaire et dynamique » (Levy, 2005, cité par Dulaurans, 2015).

Cette dimension appliquée inscrit le chercheur dans une démarche de *recherche-action* ; une démarche qui se caractérise par le fait qu'elle a été conçue dès ses origines [78] « comme obligeant le chercheur à se concentrer simultanément sur l'action et sur la production de connaissances scientifiques », comme le relève Jouison-Laffitte (2009). Cette dernière propose une définition issue de différents articles de référence :

73. Comme nous le détaillons au cours du chapitre 3, la situation a toutefois été amenée à évoluer avec l'intégration progressive de l'institution aux Archives de l'État.

74. À l'exception peut-être de questions liées plus spécifiquement à d'importants volumes de données, qui pourraient nécessiter des vérifications ultérieures.

75. Le programme BRAIN (Belgian Research Action through Interdisciplinary Networks) est un « programme-cadre de recherche récurrent qui, au travers de projets de recherche fondés sur l'excellence scientifique et l'ancrage européen et international, permet à la fois de rencontrer les besoins de connaissance des départements fédéraux et de soutenir le potentiel scientifique des Établissements scientifiques fédéraux » (Politique scientifique fédérale, 2020).

76. En l'occurence, une répartition égale (50-50) de notre temps de travail est prévue entre le Centre de recherche universitaire auquel nous sommes rattachée et l'établissement partenaire

77. En l'occurence, le CegeSoma, déjà évoqué ci-dessus, et qui sera décrit de façon plus exhaustive au cours du chapitre 3.

78. Elle a commencé à être utilisée dès les années 1950 en sciences sociales (Ottosson, 2003).

> La recherche-action est un processus ; c'est une démarche de recherche visant à résoudre les problèmes concrets en situation ; elle est mise en œuvre par une collaboration entre les chercheurs et les acteurs de l'entreprise [79] (l'implication des acteurs [...] se situe à des niveaux différents selon la conception de la recherche-action retenue par les auteurs) ; son objectif est de produire des connaissances scientifiques sur les situations étudiées. (Jouison-Laffitte, 2009)

Une telle démarche se distingue par le fait qu'elle permet un accès privilégié à des informations approfondies et de qualité qui ne pourraient pas être obtenues à l'aide de méthodes classiques (Ottosson, 2003), mais également par le fait qu'elle place le chercheur dans une position particulière : il devient un acteur central du changement opéré ou espéré [80] (Selener, 1997; Soulé, 2007). Mener à bien une telle démarche est néanmoins complexe et exigeant [81] : cela sous-entend un investissement important du chercheur qui doit assumer différents rôles et responsabilités [82] et posséder de multiples compétences : « good emotional skills, appropriate experience and knowledge, and good personal skills for the work » (Ottosson, 2003), de manière notamment à pouvoir détecter « not merely the objective problems, but also how the community evaluates its problems » (Chein *et al.*, 1948, p. 44).

Si l'implication du chercheur conditionne la qualité d'une recherche-action [83], elle en représente également l'un des principaux défis. D'une part, parce qu'elle renforce le tiraillement dans lequel se trouve le chercheur, entre la réalisation d'un travail scientifique et la satisfaction des besoins du terrain (Rapoport, 1970, cité par Jouison-Laffitte, 2009), d'autre part parce qu'elle renvoie à la question de la distanciation critique du chercheur telle que soulevée par Plane (1998).

79. Dans notre cas, il s'agit bien évidemment d'une institution publique et non d'une entreprise.

80. Comme l'ont démontré Roy et Prévost, si la recherche-action possède des points communs avec des activités de consultance, elle s'en distingue fondamentalement : voir les tableaux comparatifs qu'ils proposent : Roy et Prévost (2013, pp. 142-144).

81. « More demanding and complex than classical research », relève (Ottosson, 2003).

82. « The action researcher plays a combination of roles which include coordination of the project, assistance in diagnosing and understanding the problem, provision of technical expertise in solving the problem, and participation in the use of research results » (Selener, 1997, p. 67).

83. « The research will always gain a better understanding of his/her observations when there is maximum mental and minimal geographical proximity between the researcher and the studied objectif. In addition, there will be a deeper dimension when he/she is deeply involved in the process compared with if he/she participates occasionally » (Ottosson, 2003, p. 93).

Nous nous inscrivons dans la lignée de Ottosson, qui explique [84] qu'en acceptant ce nouveau paradigme [85] impliquant que « true objectivity does not exist, only relative objectivity, and the subjectivity of each individual », le chercheur investi dans une démarche de recherche-action « does not violate but obey modern scientific thinking when he/she is inside the studied process and is completely involved it » (Ottosson, 2003, p. 92). Il met toutefois en garde contre le fait qu'un fort engagement mental du chercheur peut l'amener à perdre sa vision d'ensemble, trop accaparé par les problèmes et les détails des processus qu'il gère. Pour éviter cela et préserver une évaluation scientifique des résultats, Ottosson recommande des contacts avec un environnement scientifique permettant de fréquentes discussions (Ottosson, 2003, p. 92), ainsi qu'une maturité suffisante, sachant que « the researcher has to come close but stay at a distance at the same time » (Ottosson, 2003, p. 93).

La question de l'implication du chercheur est particulièrement prégnante dans le cadre de thèses de doctorat issues de collaborations entre université et milieux institutionnels, étant donné qu'elles requièrent un compromis continuel entre exigences académiques et attentes professionelles. Dulaurans (2015) recense ainsi les obstacles qu'un tel dispositif engendre, de la difficulté de distanciation critique, à l'influence du mode de financement, en passant par les questions d'identité et de légitimité du doctorant. Hellec (2014) cite également la difficulté de concilier des temporalités différentes voire contradictoires, ainsi que les jeux de pouvoir dans lequel peut se retrouver entraîné le doctorant, tandis que Rouchi (2017) met en évidence des intérêts divergents, notamment en ce qui concerne l'accès aux données ou la publication des résultats. Enfin, Landon (2015) évoque la difficulté d'une posture double, par exemple lorsque le chercheur est chargé à la fois de promouvoir un projet et d'en proposer une approche critique.

Pour pallier de telles difficultés, Rouchi (2017) affirme l'importance d'adopter un point de vue réflexif et de chercher à transformer les contraintes en opportunités. Dans la même optique, Foli et Dulaurans remarquent que « s'accrocher à la finalité surplombante de la thèse permet de transformer les difficultés en expérience heuristique fertilisant finalement les travaux de recherche » (Foli et Dulaurans, 2013), tandis que Hellec note que « la recherche se nourrit tout autant de ce que l'on observe que de ce que l'on vit dans l'entreprise » et que « toutes les incompréhensions et toutes les tensions qui émaillent la collaboration avec l'entreprise sont signifiantes de la

84. Dans le cadre de *participation action research* pratiquée dans le domaine du management.

85. Il s'agit du *Quantum paradigm*, selon lequel « the researcher always influences the studied object through the tools used, no matter what tools or methods are used » (Ottosson, 2003, p. 88).

culture propre de celle-ci et de ses modes de régulation » (Hellec, 2014). Enfin, Dulaurans met en exergue le fait que le chercheur « doit alterner des phases d'intériorité et d'extériorité dans son approche du terrain et articuler sa démarche entre engagement et distanciation » (Dulaurans, 2015).

Les mêmes motifs ont émaillé le processus de recherche-action mené dans le cadre de notre étude de cas : nous avons été appelée à faire face à des dilemmes et tiraillements similaires, comme par exemple le fait d'être encouragée à travailler sur des éléments suffisamment novateurs pour être dignes d'être publiés dans des revues scientifiques tout en étant appelée à effectuer du travail de nettoyage de données afin de répondre à certains besoins opérationnels ; ou encore de nous retrouver en quelque sorte *juge et partie* du prototype élaboré dans le cadre de nos travaux [86].

Cependant, il est clair que notre statut de membre du personnel du Cege-Soma à part entière, avec tout ce que cela implique comme droits et obligations, nous a permis d'accéder à une compréhension beaucoup plus fine des enjeux et des dynamiques à l'œuvre, que ce soit dans le cadre de réunions ou de communications officielles ou lors de discussions de couloir plus informelles. Ainsi, en ayant par exemple eu l'occasion d'assister à une journée de formation interne [87] animée par un intervenant externe sur le thème des *Linked Open Data*, nous avons pu prendre conscience lors d'échanges d'une certaine méconnaissance de la part de certains membres du personnel, mais aussi de réserves et de résistance au changement s'étendant bien au-delà des propositions issues de notre travail de recherche [88].

Par ailleurs, le temps passé au sein de l'institution nous a permis de tisser des relations interpersonnelles favorisant les échanges, mais également la collaboration, au-delà des divergences de points de vue découlant de nos disciplines respectives. En effet, bien que cela n'ait pas constitué une fin en soi, cette thèse s'appuie sur une approche interdisciplinaire. Ainsi, des pratiques collaboratives ont par exemple été initiées avec des historiens et des informaticiens [89] dans le cadre de la modélisation des données destinées à être accueillies dans la Wikibase. Comme le relèvent Ravalet *et al.* au sujet d'une collaboration entre chercheurs en sciences humaines et sociales et des chercheurs en informatique, cela nécessite notamment :

86. En étant à la fois à l'origine de sa conception, poussée à en promouvoir les mérites auprès de collègues et en même temps engagée et résolue à en analyser les possibilités et les limites dans le cadre de cette thèse.

87. Destinée à une petite équipe du personnel des Archives de l'État.

88. Ainsi, nous avons par exemple pu constater certains blocages à l'œuvre lors de l'élaboration collective d'un projet fictif qui visait à appliquer les concepts théoriques abordés en première partie de formation.

89. Alors que nous sommes diplômée en communication et Gestion culturelle et que notre thèse s'inscrit dans le champ disciplinaire des Sciences et Technologies de l'Information et de la Communication.

> Un « travail d'articulation » (Strauss, 1988, cité par Ravalet *et al.*, 2017), c'est-à-dire des opérations de coordination afin de faire converger des points de vue différents, et des investissements en temps parfois coûteux. Or, la finalité de ces investissements diffère justement selon les acteurs. (Ravalet *et al.*, 2017)

Dans notre cas, si ce processus a été indéniablement enrichissant et utile pour dépasser une vision naïve et simpliste des données concernées et tendre vers une version plus éclairée et nuancée (nous y reviendrons au cours du chapitre 4, dédié à la modélisation des données), nous regrettons de ne pas avoir pu y consacrer plus de temps et qu'il ait fallu attendre les derniers mois du projet pour que la collaboration soit pleinement effective.

Enfin, il convient de signaler que ce terrain en cachait un autre, dans la mesure où notre recherche-action s'est étendue au territoire de MediaWiki et plus précisément à l'écosystème Wikidata/Wikibase, l'élaboration de notre prototype d'instance Wikibase s'avérant difficilement réalisable sans une compréhension approfondie des codes et pratiques de ces communautés. Si cela a accentué la dimension interdisciplinaire de notre thèse [90], nous amenant à compléter notre bagage académique à l'aide de nouveaux savoirs et savoir-faire, cette implication dans la communauté Wikibase a également accentué notre investissement personnel, amplifiant la difficulté de prise de recul et la surchage informationnelle.

Pour contrebalancer cette proximité avec le terrain et favoriser la plus grande distanciation critique, nous avons mis en place différentes mesures *préventives* [91] telles que suggérées par Ottosson (2003). De plus, pour préserver un équilibre *intérieur-extérieur*, nous avons veillé à maintenir un équilibre entre les temps dédiés à la recherche en tant que telle et les temps dédiés à l'*action*, de manière à ce qu'ils puissent se compléter mutuellement. En effet, comme le précise Bataille :

> On peut dire que la R.-A. [recherche-action] n'est ni de la recherche, ni de l'action, ni l'intersection des deux, ni l'entre-deux, mais la boucle récursive entre recherche et action : se situer dans la complexité c'est d'abord se situer dans cette boucle et non dans l'un ou l'autre des termes qu'elle boucle. (Bataille, 1983, p. 33, cité par Adamczewski, 1988)

90. Qui s'avère fréquente dans le contexte des sciences humaines et sociales, marqué par une porosité des disciplines et par la nécessité d'acquérir de nouvelles compétences dans le cadre de la maîtrise du numérique (Henda, 2018).

91. Comme la fréquentation régulière et continue de notre centre de recherche [92], les échanges constants avec notre directeur de recherche et nos pairs, les rencontres avec le comité d'accompagnement de notre thèse, l'organisation de réunions de *concertation* unissant nos promoteurs tant académique qu'institutionnel, ou encore la documentation régulière de notre travail, de nos observations et de nos questionnements au quotidien.

Ce processus de recherche itératif, également évoqué par Jouison-Laffitte (2009), est au cœur de notre pratique. En effet, notre première hypothèse de recherche partant des besoins du terrain s'est précisée et affinée au contact de ressources externes et au fur et à mesure de l'élaboration de notre cadre théorique. Cette dynamique s'est maintenue tout au cours de l'avancée de nos recherches, le terrain venant nuancer la théorie, et inversement, la théorie et les évolutions des contextes normatifs et technologiques [93] venant nourrir observations et expérimentations.

Plus concrètement, comme souligné précédemment, notre implication au sein de l'institution nous a permis d'obtenir un accès privilégié à des informations qui auraient difficilement pu être obtenues à l'aide de méthodes classiques. Nous avons pu accéder à tout un ensemble de sources primaires telles que les différentes bases de données de l'institution – notamment la principale, Pallas, utilisée depuis les années 1990 et contenant les descriptions de l'ensemble des collections du Centre –, de nombreux rapports, notes de vision et fichiers de documentation, les enregistrements des interviews du personnel de l'institution réalisés dans le cadre du projet MADDLAIN (Hungenaert, 2016), les procès-verbaux des réunions de l'équipe scientifique, mais également les fichiers présents sur un serveur partagé avec les Archives de l'État, ainsi que des informations de première main obtenues au cours de réunions de travail ou d'échanges informels.

En ce qui concerne les ressources externes, nos lectures se répartissent en deux ensembles : d'une part, les publications liées à notre cadre théorique, mobilisées au cours des pages précédentes et sur lesquelles nous ne nous attarderons pas ici, et, d'autre part, diverses ressources destinées à dresser un *état de l'art*. Les prochains paragraphes reviennent sur la constitution de ce corpus et ses spécificités.

Tout d'abord, précisons que, sauf exception [94], les ressources consultées ont été publiées en français, en anglais ou plus occasionnellement en allemand ou néerlandais, à une date antérieure au 1er juin 2020. Au-delà de ces critères purement pragmatiques, nous devons souligner le caractère extrêmement hétérogène des sources consultées, qui a d'ailleurs pour effet l'utilisation de l'expression *état de l'art* plutôt que *revue de la littérature scientifique*.

En effet, il est clair qu'en raison du caractère très contemporain de notre sujet d'étude [95] et des délais courant entre la rédaction d'un article, sa pre-

93. Qu'il s'agisse du développement de la nouvelle norme de description archivistique *(Records in Context)* ou du développement du logiciel de gestion d'entités liées *Wikibase*.
94. Comme Zhou *et al.* (2020), dont la publication remonte au mois d'août 2020.
95. Le présent en train de se dérouler, en quelque sorte.

mière soumission à une revue scientifique et sa publication effective[96], le corpus de sources ayant pu être mobilisées a des allures parfois peu orthodoxes. Certains passages de cette thèse sont ainsi basés sur des publications en ligne ne répondant pas toujours aux exigences de la littérature scientifique *classique*[97], voire parfois sur des sources beaucoup plus informelles telles que des extraits de *mailing lists*, des rapports de bugs, des tweets ou des échanges de messages écrits sur un groupe de conversation Telegram.

La difficulté méthodologique s'est située ici principalement au niveau de la façon d'aborder et de citer de telles sources, tout en préservant le caractère scientifique de ce travail. À cette fin, nous avons opté pour une mise en page différente lorsqu'il s'agissait de citations issus de propos d'utilisateurs ayant été publiés dans le contexte de Twitter, comme précisé en préambule. Aussi imparfaite soit-elle, cette méthode permet de fournir une certaine indication au lecteur sur la nature des sources mobilisées et leur contexte de publication.

Par ailleurs, notons que si le recours à des sources si hétérogènes pourrait apparaître comme un cas particulier difficilement évitable dans le cadre d'une thèse traitant du Web de données, il semble s'agir plus largement d'une nouvelle donne avec laquelle il faudra composer à l'avenir. Ainsi, Moirand (2004) explique aux sujets des corpus médiatiques qu'ils se caractérisent par une hétérogénéité multiforme : sémiotique, textuelle, et énonciative[98], tandis que Barats (2013) attire l'attention sur l'hétérogénéité des données provenant du web[99] et la nécessité de construire de nouvelles méthodes s'adaptant à leurs particularités.

96. Björk et Solomon ont observé à l'aide d'un échantillon de 2 700 articles issus de 135 journaux que les délais de publication et de révision (à titre indicatif, les différentes étapes de ce processus sont résumées ici par l'éditeur de revues scientifiques Wiley : https://authorservices.wiley.com/Reviewers/journal-reviewers/what-is-peer-review/the-peer-review-process.html) varient significativement selon les disciplines : « The shortest overall delays occur in science technology and medical (STM) fields and the longest in social science, arts/humanities and business/economics. Business/economics with a delay of 18 months took twice as long as chemistry with a 9 month average delay » (Björk et Solomon, 2013).

97. Telles que la validation des textes par un comité de lecture, « most often thought of as the quintessence of academic quality control » (Rigby *et al.*, 2018, p. 1088).

98. Et souligne ensuite : « Tout ensemble de textes ou de documents médiatiques recueillis constitue une somme d'occurences d'unités discursives correspondant à des pratiques langagières appartenant elles-mêmes à des séries génériques, ou dépendant de conditions de production, différentes : l'éditorial ne se confond pas avec la revue de presse, l'article d'information du journaliste scientifique avec la chronique, l'encadré explicatif avec le dessin de presse ; l'article écrit par un collaborateur régulier du journal connaît des contraintes qui ne sont pas les mêmes que celui d'un scripteur occassionnel, un spécialiste à qui l'on demande une *expertise* ou qui envoie *spontanément* son point de vue ; le correspondant régional n'a pas le même *point de vue* que l'envoyé spécial. Mais toute unité discursive peut s'inscrire dans plusieurs séries ou regroupement différents [...] » (Moirand, 2004, p. 71-72).

99. Dans un article co-écrit avec Leblanc et Fiala, elle souligne que tout corpus web « doit être regardé [...] comme un état de données provisoire, évolutif, prélevé au sein d'une archive vivante, gardant une marge d'incertitude non négligeable » (Barats *et al.*, 2013, p. 105).

Outre cette forte hétérogénéïté, l'usage de données issues du Web soulève également la question de la représentativité. Si, comme le concèdent Boullier et Lohard dans le cadre d'un projet de *sentiment analysis*, il faut renoncer à la représentativité et à l'exhaustivité avec des données issues du web, nous avons néanmoins mis en place des mesures de veille pour tendre vers l'exhaustivité. Les données recueillies permettent ainsi de dresser « une sorte de cartographie de *ce qui se dit*, dans un périmètre donné, à propos d'une ou de plusieurs thématiques » (Ravalet *et al.*, 2017). Cette veille s'accompagne également d'une démarche moins protocolaire, basée sur la sérendipité, et permettant d'étendre le champ des découvertes. En effet, comme le souligne Bevilacqua :

> Il est légitime d'envisager la démarche de l'analyste comme un mouvement de quête des traces d'inscription des usagers, et que, ce faisant, il *circule* en surface, il établit son parcours, il s'arrête là où il trouve des indices pertinents, il fouille les hypertextes, enfin, il construit son réseau de relations signifiantes lui permettant de faire ressortir le sens dont il a besoin pour expliquer les phénomènes qu'il aura ciblés davantage dans son projet de recherche. (Bevilacqua, 2016, pp. 91-92)

Enfin, il est clair que le caractère extrêmement mouvant de notre sujet d'étude, déjà évoqué au cours des paragraphes précédents, a représenté un défi conséquent. Comme l'a souligné Chartier (2013) dans le cadre de l'étude contemporaine de l'actualité dans la presse – et plus précisément de l'image de l'Islande pendant la crise économique –, la contemporanéité de l'objet étudié a une incidence sur la perspective de l'analyse. Il relate ainsi l'une des difficultés propres à l'étude d'un objet encore en constitution :

> Le temps passait pourtant et la crise se poursuivait et elle se poursuit toujours. Il fallait donc, pendant l'analyse et la rédaction, trouver une cohérence à l'objet étudié en fonction d'une chronologie fermée (l'année 2008), tout en restant sensible et attentif aux événements qui suivaient et qui influaient sur l'interprétation du corpus à l'étude selon le principe des synthèses successives dévoilé par Wolfgang Iser pour la lecture, c'est-à-dire qu'un fait nouveau peut rendre caduque l'interprétation des discours préalables. (Chartier, 2013)

Cela renvoie au problème d'acceptation de l'« impossible clôture des corpus médiatiques » et, partant, de sa « non-exhaustivité constitutive » (Moirand, 2004, p. 90). C'est une difficulté que nous avons également rencontrée. En effet, à la sensation de vertige que peut provoquer la consultation de ce

gouffre sans fond qu'est le Web [100], s'est ajouté le fait que notre travail de recherche s'inscrit dans un contexte normatif en transition dans le secteur des archives et que l'outil au cœur de notre étude de cas (Wikibase) est lui-même en développement continu. Le revers du caractère stimulant qu'offre un sujet d'étude aussi contemporain réside donc dans une tension entre l'exposition à un flux d'information continu et la volonté de tendre vers l'exhaustivité.

La seule manière de réduire cette tension semble être d'accepter cette *impossible clôture* du corpus de sources sur lequel s'appuie notre état de l'art. Il s'agit dès lors de se résigner au fait qu'un objet d'étude ultracontemporain aura pour conséquences que certains éléments seront potentiellement déjà dépassés au moment de la publication des résultats. Cette production scientifique est en quelque sorte condamnée à une inévitable obsolescence, qui n'est pas sans faire écho à la dimension de *broken-world* – évoquée dans notre cadre théorique – et soulève la question de la maintenance d'un contenu mouvant : ainsi, certains chapitres déjà rédigés sont susceptibles de ne déjà plus être tout à fait exacts et complets au moment de clôturer la rédaction [101]. Afin de lutter contre ce réflexe de vouloir sans cesse *réparer* cela, notre stratégie a donc consisté à délimiter un cadre temporel fermé à l'aide d'une borne temporelle (le 31 mai 2020) au-delà de laquelle nous n'inclurions plus de nouvelles publications.

Pour conclure, il est clair que notre position à l'intersection entre la recherche et l'action, entre les concepts et la technique, entre l'archivistique, et l'informatique, couplée au caractère très actuel de notre objet d'étude, a suscité différents défis d'ordre méthodologique. Cependant, comme Souchier l'a mis en exergue, « il n'y a pas de transformation technologique qui ne soit accompagnée d'une transformation des *modes de faire* et par là même des *modes de pensée* » (Souchier, 1996, p. 106, cité par Bourdeloie, 2014). Or, c'est précisément grâce à cette position, un peu en dehors, un peu en dedans de ces différentes sphères, que nous avons pu tenter d'appréhender certaines de ces transformations à travers la rédaction de cette thèse.

100. L'écrivain français Michel Déon écrivait dans ses Lettres de château que « La Connaissance est un gouffre sans fond. On en revient, quand on revient, illuminé ou fou » ; la croissance exponentielle de la publication d'information sur le Web ne semble pouvoir que renforcer ce phénomène.

101. Nous pensons par exemple à la question de la synchronisation entre des données issues de différentes instances Wikibase, abordée au cours de la seconde partie de cette thèse : une question que WMDE a annoncé vouloir traiter au cours de la seconde partie de l'année 2020 (Wikidata, 2020b) qui coïncide avec la période de fin de rédaction de cette thèse.

Première partie

État de l’art

Introduction

Cette thèse est divisée en deux parties principales. Cette première partie est dédiée à l'état de l'art, qui est composé de deux chapitres.

Le premier chapitre, intitulé *Vers des entités Personne* esquisse l'évolution de trois normes internationales de description archivistique : ISAD(G) [102], ISAAR (CPF) [103] et RiC [104]. Il aborde leur émergence, leur réception et leur implémentation, en mettant l'accent sur la place accordée aux personnes physiques. Il est émaillé d'exemples permettant d'aborder la façon dont les données d'autorité liées aux personnes vont progressivement être envisagées sous la forme des entités faisant l'objet de nouvelles formes de publication et de valorisation dans le contexte du Web de données.

Le deuxième chapitre, intitulé *Gestion des données structurées* s'intéresse à la façon dont les institutions peuvent envisager la gestion de leurs données d'autorité dans un contexte où ces dernières sont destinées à être structurées et sémantisées de manière à pouvoir être lisibles par des machines. Ce chapitre se concentre autour du logiciel libre Wikibase : la première section expose l'origine et l'évolution de ce logiciel, tandis que la seconde section vise à aborder les possibilités et limites de Wikibase pour les institutions du patrimoine culturel, en se concentrant sur la dimension de maintenance.

102. ISAD(G) pour International Standard Archival Description-General.
103. ISAAR (CPF) pour International Standard Archival Authority Records for Corporate Bodies, Persons, and Families.
104. RiC pour Records in Contexts.

1 | Vers des entités Personne

Introduction

Ce premier chapitre se concentre sur l'évolution des données d'autorité dans le secteur des archives. Il vise à exposer la façon dont la description des personnes physiques [1] liées à des fonds d'archives a évolué, d'une simple mention de nom de personne dans un inventaire, à une entité reliée à des milliers d'autres au cœur du Web de données.

Ce chapitre cherche à identifier les principales étapes et les enjeux concernant le traitement des noms de personnes, tout en tenant compte du contexte plus global, à savoir : les normes internationales de description archivistique [2]. Leur apparition, jugée tardive (Sibille, 2012a), remonte aux années 1990 [3]. Ces normes sont décrites par le Comité des normes et bonnes pratiques (CBPS) [4] du Conseil International des Archives (ICA) [5] comme « des

1. Précisons ici la raison de l'utilisation des termes *personne physique*. Si, dans l'usage courant, il est aisément compris que le terme *personne* désigne un individu, il convient dans un contexte de gestion documentaire de spécifier s'il s'agit de personnes physiques ou morales. Ces termes juridiques permettent en effet de distinguer la personne physique, à savoir la personne prise en tant qu'individu – un être humain doté de la capacité juridique et titulaire de droits et de devoirs – (BeCompta, 2020b) de la personne morale, qui désigne une entité juridique (un établissement ou un groupement de personnes ou de biens) à laquelle est conférée la personnalité juridique (BeCompta, 2020a). Cette précision nous semble d'autant plus nécessaire que les services d'archives ne collectent pas seulement des fonds issus de personnes [physiques], mais également de collectivités (entreprises, associations et autres regroupements).

2. Les normes existantes sont au nombre de quatre, à savoir : la Norme générale internationale de description archivistique (ISAD(G)) ; la Norme internationale sur les notices d'autorité utilisées pour les archives relatives aux collectivités, aux personnes et aux familles (ISAAR(CPF)) ; la Norme internationale pour la description des fonctions (ISDF) ; la Norme internationale pour la description des institutions de conservation des archives (ISDIAH). Une nouvelle norme, Records in Contexts (RiC), est actuellement en cours d'élaboration, comme nous le verrons au cours des pages suivantes. Pour plus de détails, voir : https://www.ica.org/fr/ressources-publiques/normes.

3. Bien que le premier énoncé de principes de l'ICA, duquel vont dériver ces normes, ait été initié en 1989 déjà (avant d'aboutir et d'être publié en 1992) (International Council on Archives, 1992).

4. Qui fait suite, depuis 2004, au Comité sur les normes de description (CDS).

5. Fondée en 1948, cette organisation neutre non-gouvernementale a pour objectif « la gestion efficace des archives et de la conservation, le traitement et l'utilisation du patrimoine

lignes directrices reflétant un consensus d'associations commerciales ou d'organismes industriels, professionnels ou gouvernementaux, reconnus nationalement ou internationalement, sur des produits, des pratiques ou des opérations » (International Council on Archives, 2016b). Elles sont destinées à fournir aux professionnels des archives des modèles de représentation des entités archivistiques servant également à la conception de systèmes informatiques. Elles s'inscrivent dans le respect des principes de base de l'archivistique – comme le principe de la provenance et le principe du respect des fonds – et visent à « surmonter de manière constructive les difficultés découlant des différentes traditions archivistiques [6] » (Sibille, 2012a, p. 88).

Bien que ce chapitre n'ait pas pour vocation de présenter une vue exhaustive de l'évolution des normes [7], sa structure repose toutefois sur l'émergence de trois d'entre elles : ISAD(G), ISAAR (CPF), ainsi que RiC – actuellement encore en cours d'élaboration –, qui seront chacune abordées au sein d'une section spécifique, sous l'angle des entités Personne. Ces trois sections sont construites sur une structure similaire : une première sous-section se penche brièvement sur le contexte d'apparition de ces normes et les besoins qu'elles sont destinées à couvrir ; la deuxième sous-section porte sur leur chronologie, leur contenu et leur réception ; enfin, une troisième sous-section est dédiée à leur implémentation concrète et est illustrée à l'aide d'exemples.

1.1 Des mentions

1.1.1 Contexte

Comme l'explique Stibbe, la nécessité de standardisation des descriptions archivistiques est apparue dans les années 1980 avec l'émergence des premiers projets d'informatisation dans le domaine des archives, en particulier au Canada et aux États-Unis. Ces expérimentations ont en effet mis en évidence une absence générale de cohérence dans les pratiques descriptives (Stibbe, 1998, p. 132). Comme le relève Sibille :

> Jusque-là, une opinion couramment répandue chez les archivistes était que chaque fonds avait son propre classement, ses caractéristiques propres et qu'il était le résultat d'un processus historique unique. Les instruments de recherche étaient donc cen-

archivistique mondial ; à ce titre, il représente les professionnels des archives du monde entier » (International Council on Archives, 2020).

6. Bien qu'il ait été souligné qu'elles ne reflètent pas l'ensemble des traditions archivistiques, mais avant tout des perspectives européo-australo-nord-américaines (Sibille, 2012a, p. 89).

7. Pour ce faire, nous renvoyons plutôt vers des publications spécialisées (Sibille, 2012a ; Davies, 2017 ; Janssens de Bisthoven, 2020 ; Popovici, 2019).

> sés présenter des caractéristiques très spécifiques, rendant impossibles toute comparaison. (Sibille, 2012a, p. 78)

Par ailleurs, une autre des motivations en faveur d'une normalisation des descriptions d'archives résidait dans la volonté de pouvoir échanger des informations entre dépôts d'archives, mais également à un niveau plus international :

> The scattering of archival fonds among multiple archives or among countries, etc. was also seen as a reason for standardising descriptive information so as to make re-describing material belonging to the same fonds, or copies of these, no longer necessary and to enable collocation, i.e., the bringing together, of parts of fonds of the same provenance possible in union listings or finding aids, such as institutional, regional or national finding aids. This latter issue is becoming more relevant and pressing when archival repositories are making their holdings accessible on-line on the Internet in the form of descriptions representing those holdings. (Stibbe, 1998, p. 133)

C'est pour répondre à ces besoins, en partant des principes fondamentaux du respect des fonds et de la description à plusieurs niveaux, mais également des outils normatifs existants (Sibille, 2012b), que la première norme internationale de description archivistique – ISAD(G) – va voir le jour, devenant ainsi « l'un des jalons les plus importants de l'histoire de la science archivistique, qui a permis de résumer [et] rassembler, à l'échelle internationale, les expériences professionnelles de nombreux pays » (Popovici, 2019, p. 22).

1.1.2 ISAD(G)

Si la première version de la Norme générale et internationale de description archivistique, généralement évoquée à l'aide de l'acronyme ISAD(G) – pour International Standard Archival Description-General –, est publiée en 1994, les prémices de règles communes remontent à la fin des années 1980 déjà. Nous reprenons ci-dessous les grandes étapes de ce développement :

- 1988 : un groupe d'experts réunis à Ottawa [8] invite l'ICA à former un groupe de travail destiné à développer des standards internationaux pour la description archivistique (International Council on Archives, 1992)
- 1990 : l'ICA crée l'Ad Hoc Commission on Description Standards (ICA/DDS) (International Council on Archives, 1992)
- 1992 : l'ICA/DDS approuve une version révisée du document *Statement of Principles Regarding Archival Description*, ainsi que la version

8. À l'initiative de l'ICA et des Archives nationales du Canada.

de travail[9] de l'*International Standard Archival Description-General* (ISAD(G), qu'il soumet aux commentaires (International Council on Archives, 1992)
— 1994 : l'ICA/DDS publie la première version de la norme ISAD(G) (Sibille, 2012a)
— 1996 : l'ICA/DDS devient un comité permanent de l'ICA : le *Committee on Descriptive Standards* (ICA/CDS) (International Council on Archives, 2000)
— 1998 : l'ICA/CDS soumet aux commentaires une version de travail de la deuxième édition de la norme ISAD(G) (International Council on Archives, 2000)
— 2000 : l'ICA/DDS publie la deuxième édition de la norme ISAD(G) (International Council on Archives, 2000)

Cette norme ayant pour objectif de faciliter l'accès aux documents d'archives est destinée à être utilisée soit en relation avec les normes nationales existantes, soit comme point de départ pour le développement de normes nationales (International Council on Archives, 2000). Elle est constituée de 26 éléments – répartis en sept zones d'information (identification, contexte, contenu, conditions d'accès, sources complémentaires, notes, contrôle de la description) – dont seuls six sont jugés essentiels pour permettre l'échange international de descriptions archivistiques : la référence ; l'intitulé ; le producteur ; la ou les dates ; l'importance matérielle de l'unité de description ; le niveau de description.

En ce qui concerne les mentions de noms de personnes, il est intéressant de noter que dès la première édition de la norme, le concept d'*access point*[10], par ailleurs déjà présent dans le *Statement of Principles Regarding Archival Description* (International Council on Archives, 1992), est évoqué. La norme suggère d'ailleurs d'accroître la valeur de ces points d'accès en passant par le contrôle d'autorité[11] et mentionne son intention de développer des normes internationales à ce sujet (Gueguen *et al.*, 2013). Précisons encore que les vocabulaires et conventions associés à ces points d'accès sont destinés à être élaborés à l'échelle nationale ou internationale, séparément pour chaque langue.

Apparue voilà plus de 25 ans, cette norme a influencé de façon significative les pratiques internationales de description archivistique. Le Groupe d'experts sur la description archivistique (EGAD) de l'ICA estime ainsi que :

9. Voir : https://cool.culturalheritage.org/lex/icoh.html.
10. Définis comme « [un] nom, mot-clé, entrée d'index, etc., permettant de rechercher et de retrouver une description. ».
11. C'est-à-dire le « contrôle des formes normalisées des termes, y compris des noms propres (noms de personnes physiques ou morales, ou noms géographiques), utilisés comme points d'accès ».

> Even if new types of finding aids and archival descriptive systems based on the four ICA standards have been developed in the past ten years, the predominant model of archival description remains the fonds-down hierarchical description prescribed in ISAD(G). For a variety of reasons, it is likely to remain the prevailing approach to archival description for the near future : it addresses the traditional understanding of the Principle of Provenance ; it is well understood by the community ; a variety of existing methods and systems exist to facilitate creation, maintenance, and publication ; and finally, it is a relatively economic approach to an exceptionally complex, labour-intensive challenge. (International Council on Archives Expert Group on Archival Description, 2016, p. 8)

Un format d'échange est apparenté à la norme ISAD(G) : l'Encoded Archival Description (EAD). Ce standard d'encodage reposant sur le langage XML fait l'objet de la section suivante.

1.1.3 EAD

Le développement de l'Encoded Archival Description (EAD) [12], ce standard d'encodage des instruments de recherche archivistiques basé sur le langage XML, a débuté en 1993 à l'Université de Californie à Berkeley. La première version officielle a été publiée en 1998, suivie par une deuxième et une troisième version, respectivement en 2002 et 2015 [13]. Il faut noter que le standard, basé sur le développement d'une DTD (Document Type Definition) dont la Library of Congress assure la maintenance, a été développé en parallèle de l'ISAD(G). Cela signifie que, bien qu'il reflète le même type de compréhension de la description archivistique que l'ISAD(G), des différences notables puissent être relevées (Sibille, 2012a, p. 85).

12. Les fichiers et la documentation associés aux différentes versions de l'EAD sont mis à disposition par la Library of Congress : https://www.loc.gov/ead/, de même que les archives de la *mailing list* dédiée à l'EAD et à l'EAC-CPF https://listserv.loc.gov/cgi-bin/wa?A0=EAD.

13. Pour une rétrospective revenant, 20 ans après sa création, sur l'émergence et la réception de ce standard, mais également les freins à son adoption (« lack of institutional support for new technology and resources for staff ; the time and effort it takes to encode, or convert to, EAD ; the need for knowledge and expertise to implement EAD ; sensitive content closed to the public ; a low comfort level with providing public access ; and recent establishment »), voir Eidson et Zamon (2019).

La DTD-EAD de la seconde version (EAD2002) [14] contient 146 éléments, dont huit seulement sont obligatoires (Groupe AFNOR CG46/CN357/GE3, 2009). En ce qui concerne les personnes physiques, nous pouvons relever au sein de la seconde version de l'EAD (2002) trois éléments en particulier [15] :

<persname> élément permettant l'indexation de noms de personnes dans un instrument de recherche, soit directement au sein du texte [16], soit au sein de l'élément <controleaccess> (vedettes et accès contrôlés), qui est placé à la fin de la description archivistique (<archdesc>).

<origination> élément [17] fournissant des informations sur la personne physique ou morale qui a produit, rassemblé ou constitué les unités documentaires décrites, avant leur intégration dans une institution responsable de l'accès intellectuel (Society of American Archivists, 2004, p. 178). Dans le cas d'une personne physique, cette dernière peut être indiquée à l'aide du sous-élément <persname>.

<bioghist> élément contenant un texte rédigé ou une chronologie, qui place de façon concise les documents d'archives dans leur contexte, en fournissant des informations à propos de leur(s) producteur(s). Cela inclut des informations significatives sur la vie d'un individu ou d'une famille, ou sur l'histoire administrative d'une collectivité (Society of American Archivists, 2004, p. 55). Dans le cas d'une personne

14. Bien qu'une troisième version ait été émise en 2015 par le Technical Subcommittee for Encoded Archival Description of the Society of American Archivists, nous nous référons à cette seconde version étant donné qu'elle est accompagnée d'une solide documentation, qu'elle est notamment utilisée par le CegeSoma – l'institution au cœur de l'étude de cas présentée en seconde partie de cette thèse – et que l'adoption de l'EAD3 est encore limitée (Tillman, 2016).

15. Nous pourrions également penser à <custodhist> (historique de la garde des documents) qui, combiné à l'élément <persname> (nom de personne), permet d'indiquer des collectionneurs, « c'est-à-dire des personnes ou collectivités qui ont réuni artificiellement un ensemble de documents en fonction de critères communs liés à leur contenu ou à leur support, par opposition aux fonds d'archives constitués de façon organique. » (EAD (ead-bibliotheque.fr), 2020b), lorsque la description archivistique ne concerne pas la collection dans son ensemble.

16. EAD (ead-bibliotheque.fr) (2020a) recommande « de procéder à une indexation au fil du texte chaque fois que cela est possible », cependant, comme le souligne ce manuel de la Direction des Archives de France, « cet encodage fin prend beaucoup de temps, et par conséquent pèse sur les ressources d'une institution. [...] D'autre part, un encodage trop systématique des informations de contenu risque d'induire l'utilisateur en erreur : il met en effet en valeur des termes qui n'ont pas nécessairement grande importance pour la description des documents (des noms propres, par exemple » (Groupe AFNOR CG46/CN357/GE3, 2009, p. 10).

17. Utilisé au plus haut niveau de description (<did>), mais il peut éventuellement être utilisé à un autre niveau pour préciser un sous-producteur s'il est connu – par exemple pour indiquer le nom de l'auteur d'une œuvre graphique (Groupe AFNOR CG46/CN357/GE3, 2009).

> physique, cette dernière peut être indiquée à l'aide du sous-élément <persname>. Enfin, notons que si <bioghist> est avant tout utilisé pour décrire le producteur du fonds, il peut également être utilisé à d'autres niveaux de la description (EAD (ead-bibliotheque.fr), 2020a).

Notons également que l'élément <persname> est destiné à être complémenté à l'aide d'attributs [18]. L'attribut @role – dont les valeurs sont destinées à être issues d'une liste fermée de valeurs – est par exemple utilisé pour indiquer la fonction de cette personne, comme *photographe*, *collectionneur*, *créateur* ou encore *auteur*, *sujet*, *responsable scientifique*, tandis que l'attribut @normal vise à spécifier la forme normalisée d'un nom, de façon distincte du texte destiné à être affiché. L'EAD prévoit également des renvois vers des notices d'autorité : les attributs @source et @authfilenumber permettent de « fournir les informations nécessaires à un programme ou à un logiciel pour établir des liens vers des notices d'autorité [19] » : à savoir le référentiel utilisé – stocké à l'aide de l'attribut @source – et le numéro de la notice d'autorité concernée – indiqué dans l'attribut @authfilenumber.

En ce qui concerne la troisième révision de l'EAD (EAD3), qui est toujours basée sur un encodage en XML, mais vise notamment à faciliter l'insertion de *Linked Data* directement au sein des descriptions archivistiques (Tillman, 2016), elle comporte deux principaux changements concernant le traitement des noms de personnes. D'abord, @authfilenumber a été remplacé par @identifier, tandis que @relator remplace maintenant @role ; de manière à harmoniser la terminologie au langage utilisé par les bibliothèques et les *Linked Data*. De plus, le type de données accepté est maintenant restreint au *token* – qui exclue tout saut de ligne, espace de début ou de fin ou tabulations. Enfin, l'élément <part> a été intégré, permettant par exemple de décomposer les noms en données plus finement structurées (Tillman, 2016, p. 27). Cependant, comme le relève Tillman, même la première version de l'EAD était déjà propice aux *Linked Data* :

> From its inception, EAD has contained within it elements and attributes which could be used to encode URIs for use as linked data. [...] Although early adopters may not have had linked

18. Afin d'alléger le texte, nous citons exceptionnellement dans une seule note de bas de page les deux ressources ayant été utilisées dans le cadre de la rédaction de ce paragraphe : EAD (ead-bibliotheque.fr) (2020c) ; Groupe AFNOR CG46/CN357/GE3 (2009).

19. Ces renvois permettent « de répercuter automatiquement dans les instruments de recherche les modifications apportées dans le fichier d'autorité ; d'enrichir les index au moyen des formes rejetées présentes dans les notices d'autorité ; d'établir des liens hypertextuels entre les instruments de recherche et le fichier d'autorité ; de remplir automatiquement un élément <bioghist> avec les données présentes dans le fichier d'autorité ; de rendre les instruments de recherche accessibles aux applications du Web sémantique par l'intermédiaire du fichier d'autorité, etc. » (EAD (ead-bibliotheque.fr), 2020c).

> data in mind when they used them [@authfilenumber, @source, @role], if they used these attributes when referencing authorities, they laid the foundation for their successors to transform and reuse their data as linked data. (Tillman, 2016, p. 20)

Après avoir abordé la façon dont les noms de personnes sont pris en charge, sous forme de mentions, par la norme ISAD(G) et son standard d'encodage apparenté, l'EAD, il est temps de s'intéresser à l'émergence de notices d'autorité à part entière, dans le cadre d'une nouvelle norme internationale de description archivistique.

1.2 Des notices

1.2.1 Contexte

Comme l'explique Stibbe – l'un des principaux créateurs de l'ISAD(G) : « as a result of the work on the ISAD(G) and the comments received on the draft, the commission realized that it had only half a standard », dans le sens où la norme ne couvrait pas optimalement la question des producteurs des documents (Stibbe, 1998, p. 140). C'est ainsi que va être initiée la création de la seconde norme internationale publiée par l'ICA : the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR (CPF)).

Mais la nécessité d'établir des notices d'autorité de provenance devant être tenues rigoureusement séparées de la description des archives est déjà propagée au cours des années 1980[20] par des figures comme Bearman et Lytle. Ils s'emploient à expliquer que de tels fichiers d'autorité devraient permettre d'obtenir de la cohérence dans l'usage des points d'accès concernant des noms, mais aussi d'éviter la confusion que peut générer la présence d'informations au sujet des producteurs au milieu d'informations sur les archives elles-mêmes, par exemple dans les préfaces d'inventaires. Ils affirment par ailleurs que les relations entre organismes producteurs d'archives devraient être spécifiées dans le contenu intellectuel des notices d'autorité de provenance, de manière à pouvoir être exploitées par les index lors de la recherche d'information Bearman et Lytle (1985). Tout sera formalisé avec l'apparition de la norme ISAAR (CPF), dont la première version est publiée en 1996.

1.2.2 ISAAR (CPF)

ISAAR (CPF) – la Norme internationale sur les notices d'autorité archivistiques relatives aux collectivités, aux personnes et aux familles – constitue la

20. Thibodeau souligne que l'idée avait toutefois déjà été conceptualisée par Peter Scott en 1966.

seconde norme publiée par l'ICA. Son élaboration avait toutefois déjà débuté avant que ne soit publiée la première version de la norme ISAD(G) Gueguen *et al.* (2013), comme le montrent les grandes étapes de son développement :

- 1993 : un sous-groupe de l'ICA/DDS *(Ad Hoc Commission on Description Standards)* élabore un document de travail portant sur les points d'accès et propose l'élaboration d'une norme pour les notices d'autorité archivistiques (International Council on Archives, 1996b)
- 1995 : l'ICA/DDS soumet aux commentaires une version de travail de la première édition de l'International Standard Archival Authority Records—Corporate Bodies, Persons, and Families (ISAAR (CPF)) (Thibodeau, 1995)
- 1996 : l'ICA/DDS publie la première édition de la norme ISAAR (CPF) (International Council on Archives, 2004) ; l'ICA/DDS devient un comité permanent de l'ICA, désormais nommé le *Committee on Descriptive Standards* (ICA/CDS) (International Council on Archives, 2004)
- 2000 : l'ICA/CDS soumet aux commentaires une version de travail de la deuxième édition de la norme ISAAR (CPF) (International Council on Archives, 2004)
- 2004 : l'ICA/CDS publie la deuxième édition de la norme ISAAR (CPF) (International Council on Archives, 2004)

Comme Gueguen *et al.* l'ont montré, ce concept de notices d'autorité a été calqué sur les systèmes d'autorités liées utilisés dès les années 1980 par les bibliothèques, à qui les archives empruntèrent également le principal argument en faveur du développement d'une telle approche, à savoir : les économies que cela permet (Gueguen *et al.*, 2013). Les notices d'autorité envisagées par la communauté archivistique sont toutefois appelées à se distinguer de celles des bibliothèques, comme le précise la norme ISAAR(CPF) :

> Une notice d'autorité pour les archives est semblable à une notice d'autorité pour les bibliothèques dans la mesure où, dans les deux cas, il convient de créer des points d'accès normalisés à la description. Le nom du producteur de l'unité de description est un des points d'accès les plus importants. Tout point d'accès peut comporter des qualificatifs, essentiels pour identifier l'entité ainsi désignée et permettre de distinguer sans ambiguïté différentes entités qui ont le même nom ou des noms très proches. (International Council on Archives, 2004, p. 7)

Concrètement, la norme ISAAR(CPF) vise à améliorer l'accès pour les utilisateurs et à diminuer les re-créations multiples de ressources par le partage des descriptions (International Council on Archives, 2004, p. 8). Plutôt que de proposer des règles spécifiques pour la création de formes autorisées

des noms [21], elle prescrit la nature des informations devant être inclues, qui sont regroupées en quatre zones : zone d'identification ; zone de description ; zone des relations ; zone du contrôle de la description (International Council on Archives, 2004). Par ailleurs, il faut relever que la norme met avant tout l'accent sur le producteur d'archives, bien que les notices d'autorité puissent également servir à contrôler « la forme du nom et l'identité de toute collectivité, personne ou famille citée dans tout point d'accès relié à une unité de description » (International Council on Archives, 2004, p. 8).

1.2.3 EAC-CPF

Initiée par l'Université de Yale en 2001 [22], l'Encoded Archival Context (EAC) a pour but d'offrir un format de communication pour l'échange de notices d'autorité conformes à la deuxième édition de la norme ISAAR (CPF). Ce standard XML vise à permettre la production de documents pérennes, indépendamment de tout logiciel (Sibille, 2012a). La première version du standard a été finalisée et publiée en 2010 après une longue période de test de la version beta. Elle fait actuellement l'objet de révisions, notamment dans l'optique d'une réconciliation avec l'EAD3 [23] (Arnold, 2019).

Plutôt qu'une analyse détaillée du standard, nous voulons plutôt ici illustrer la façon dont cette standardisation de la description des notices d'autorité archivistiques a ouvert la voie à de nouvelles formes d'échanges de données. Sachant que le travail de création et de gestion des fichiers d'autorité est fréquemment décrit comme l'un des plus coûteux en temps et en argent (SNAC, 2017; Byrum Jr, 2004), il n'est pas étonnant de constater que de nouvelles pratiques collaboratives aient progressivement vu le jour. Comme l'expliquent Waibel et Erway dans un article consacré aux collaborations possibles entre bibliothèques, archives et musées, les notices d'autorité peuvent être mises au service d'une vision basée sur la devise *think globally, act locally* :

> Whether describing mass-produced items or unique materials, the same basic concepts are of prime interest : places, names, dates, object types, to just name the most obvious examples. Enabling all communities with a vested interest in these concepts to contribute to an authority record would leverage the expertise of the entire LAM [Libraries, Archives, Museums] community, and

21. La norme précise que « le contenu de ces informations sera fixé par les règles ou conventions en usage dans le service qui rédige la notice d'autorité » (International Council on Archives, 2004).

22. Se référer à Society of American Archivists (2004) pour un historique détaillé.

23. Les progrès de cette révision peuvent être suivis en direct en consultant le dépôt GitHub dédié au schéma EAC : https://github.com/SAA-SDT/eac-cpf-schema/issues.

> create new economies of scale. [...] [Doing so] the core activity of cataloging [will consist] of finding the right pointer(s) to the right authority record(s), and aiding the distributed effort of updating and maintaining the authority file. (Waibel et Erway, 2009, p. 332)

Cependant, si cette pratique est devenue commune dans le monde des bibliothèques, elle semble toutefois moins intuitive dans le secteur des archives :

> Copy cataloging has no significant role to play in archives. But if descriptions of the records themselves cannot be usefully shared, are there components of the description that overlap the holdings of archives that would provide an economic motivation for archives to share the work of creating and maintaining them ? (SNAC, 2017, p. 2)

Selon les initiateurs du projet Social Network and Archival Context (SNAC) – qui sera évoqué plus longuement dans les paragraphes suivants –, c'est précisément la description contextuelle qui fournit aux archivistes un travail pouvant être efficacement exécuté de manière collaborative et permettant notamment de réaliser de nouvelles économies dans le traitement et la description des documents (SNAC, 2017).

Outre ces potentielles économies, il s'agit également de la nécessité pour les institutions de pouvoir se concentrer sur leur cœur de métier. Dans le contexte voisin des Beaux-Arts, Terry Gould explique ainsi la raison pour laquelle les pages *artistes* web des Galeries nationales d'Écosse ont été enrichies à l'aide de données disponibles en interne, mais également à l'aide de deux sources externes [24], grâce aux données liées :

> Another consideration is the wealth of information that already exists about many key artists – if you consider a significant European artist like Gauguin, there are hundreds of biographies of him that already exist, from a wealth of reputable, knowledgeable and trustworthy sources [...]. Rather than reinvent the wheel for every artist, we can use trusted sources to give our pages the necessary contextual information ; allowing our interpretation to focus more closely on the artist's processes and approaches to the specific works we have in the collection and the stories that we are best placed to tell. (Gould, 2018)

C'est dans ce contexte qu'apparaît le projet SNAC, qui se présente aux bibliothécaires et archivistes comme une plateforme « for sharing labor-intensive identity resolution and description of shared CPF [Corporate bodies, Persons, Families] entities » (SNAC, 2020a). Le projet SNAC voit le jour en 2010

24. À savoir : Wikidata et the Getty's Union List of Artist Names.

aux États-Unis, sur l'initiative de trois universités partenaires [25] et plus précisément sous la conduite de Daniel Pitti, le principal *architecte* de l'EAD et de l'EAC-CPF. Envisagé au départ comme un projet de recherche et de démonstration – financé par the U.S National Endowment for the Humanities –, il évolue en 2015 – grâce au financement additionnel de l'U.S. Institute for Museum and Library Services et l'Andrew W. Mellon Foundation – pour devenir une coopérative internationale (SNAC, 2020b) comptant aujourd'hui 42 membres (SNAC, 2020a). Le principal fruit du projet est une plateforme de découverte [26] de « persons, families, and organizations found within archival collections at cultural heritage institutions ».

Ce faisant, SNAC se donne la mission de relever un défi de longue date dans le domaine de la recherche : « discovering, locating, and using distributed historical records » (SNAC, 2020b). Notant que les données permettant d'affronter ce challenge existent déjà à travers les inventaires et autres instruments de recherche créés par les archivistes et bibliothécaires, les initiateurs du projet regrettent qu'elles soient toutefois enfouies dans différents silos et ambitionnent de fournir un accès centralisé à ces données, en séparant les descriptions des personnes, familles et collectivités des descriptions des fonds d'archives (SNAC, 2020b). Le projet apparaît donc comme un moyen de tester à l'aide de données empiriques le schéma EAC-CPF « the last essential component for developing linked archival description systems that support sharing work and data » et de montrer à la communauté archivistique « the economies of cooperative description » (SNAC, 2017, p. 6).

L'une des dimensions innovantes du projet SNAC est certainement la façon dont les descriptions EAC-CPF au cœur de la plateforme découverte – dont la figure 1.1 propose un aperçu – ont été créées : la plupart des notices ont été générées de façon automatisée à partir de données extraites de descriptions d'archives à l'aide de méthodes computationnelles [27] (SNAC, 2020c). Concrètement, environ 3,7 millions de notices [28] (concernant tant des personnes que des familles ou collectivités) ont été créées à partir de 150 000 inventaires EAD et 2,2 millions de notices sur des fonds d'archives ayant été encodées au format MARCXML [29] (Casalini *et al.*, 2018, p. 28). Après leur déduplication, ces données ont par ailleurs été réconciliées avec

25. À savoir : the University of Virginia, Institute for Advanced Technology in the Humanities ; the University of California, Berkeley School of Information ; and the University of California, California Digital Library.

26. https://snaccooperative.org/.

27. Le script permettant la conversion de l'EAD vers l'EAC est librement mis à disposition, comme le reste des scripts utilisés, voir : https://github.com/snac-cooperative/snac_ead_to_cpf.

28. En juillet 2020, la plateforme regroupe 3 735 767 millions d'entités (*identity constellations*) et 2 092 042 millions de descriptions (SNAC, 2020d).

29. Pour avoir le détail sur la provenance de ces données, voir : SNAC (2020c).

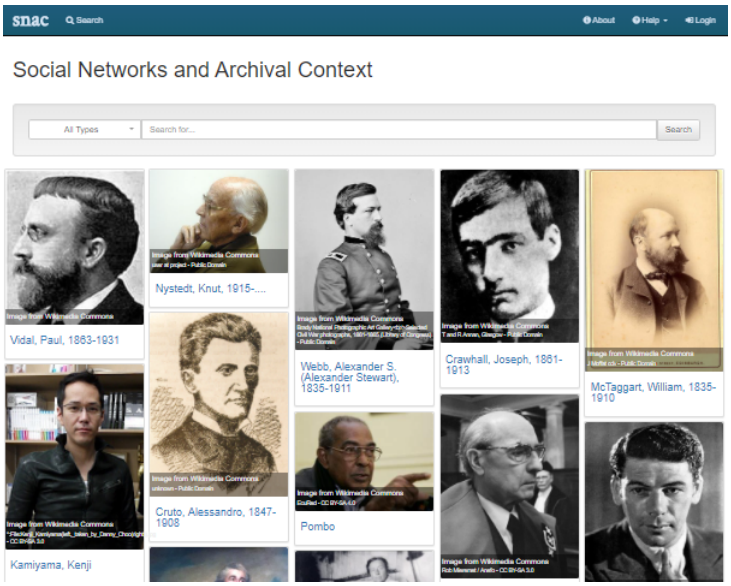


FIGURE 1.1 – Aperçu de la plateforme SNAC. Source : SNAC (https://snaccooperative.org/).

les plus de 25 millions d'entités que contient VIAF (Virtual International Authority File) [30] (SNAC, 2020c).

La figure 1.2 montre un exemple de page d'une entité Personne sur la plateforme SNAC [31], il s'agit dans ce cas-ci de l'américaine Ethel Rosenberg, qui a été condamnée à mort avec son mari Julius après avoir été acccusés d'espionnage pour le compte de l'URSS au début de la Guerre froide. La page contient des informations biographiques [32], des formes alternatives du nom [33], des liens vers des ressources externes équivalentes comme VIAF ou WorldCat Identities, la mention de l'identifiant SNAC et de l'identifiant ARK (Archival Resource Key) associés à l'entité, une liste [34] de relations [35] liant cette entité à d'autres personnes, familles ou collectivités, et la plateforme prévoit que des informations puissent également être fournies sur des places, fonctions, sujets ou occupations liées à cette personne. C'est l'onglet *resources* – visible sur la capture d'écran proposée en figure 1.2 – qui permet à l'utilisateur d'accéder aux descriptions de fonds d'archives associées à cette personne. SNAC distingue deux types de rôle pour l'entité associée aux

30. http://viaf.org/.
31. https://snaccooperative.org/view/9353978.
32. Incluant une notice biographique issue de Wikipédia en anglais, qui ne porte en l'occurence pas exclusivement sur Ethel Rosenberg, mais sur le couple qu'elle formait avec Julius Rosenberg.
33. Récupérées de VIAF, comme le montre la page *sources* : https://snaccooperative.org/sources/9353978/11342353.
34. Également disponible sous forme de data visualation.
35. Dans le cas présent, il s'agit de relations associatives (*associatedWith)*.

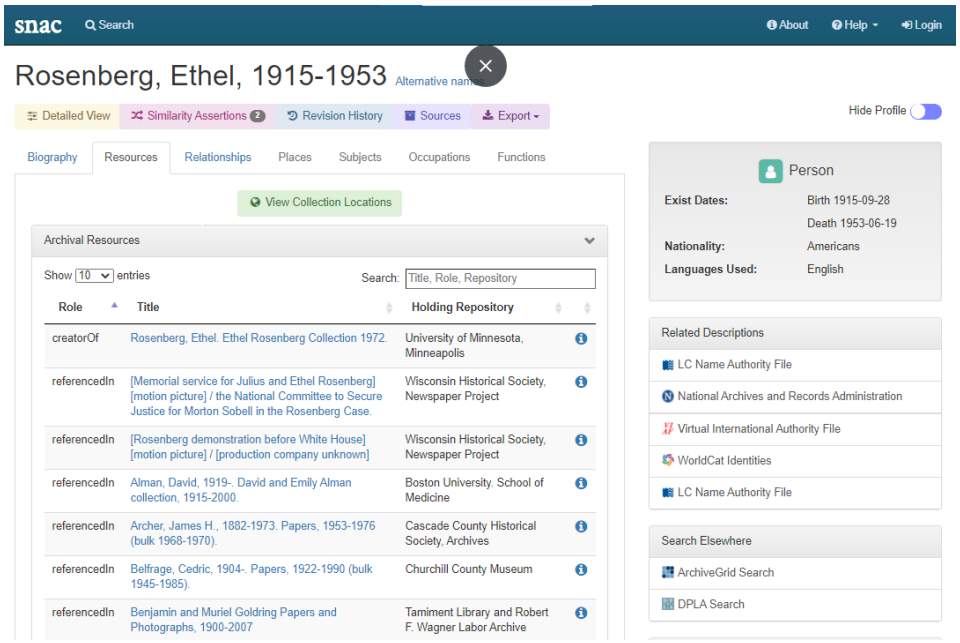


FIGURE 1.2 – Ethel Rosenberg sur la plateforme SNAC. Source : SNAC (https://snaccooperative.org/view/9353978).

fonds : ainsi, Ethel Rosenberg est listée comme producteur *(creatorOf)* d'un fonds et comme référencée *(referencedIn)* dans 93 autres fonds.

Il ne faudrait pas oublier l'onglet *Similarity Assertions*, qui constitue l'une des spécificités du projet SNAC : il intègre à l'aide du qualificatif *mayBeSameAs* [36] des liens vers d'autres entités correspondant potentiellement à la même personne. En effet, comme l'explique une ancienne page d'information du projet :

> Users will frequently find two or more SNAC identity descriptions for the same person or organization. In attempting to resolve identities, SNAC has intentionally erred on the side of not combining similar identities when there is uncertainty. The rationale for this bias is that it can be exceptionally difficult for users to identify when distinct identities have been incorrectly combined. SNAC thus attempts to avoid such combinations, even at the expense of not combining what to human users are seemingly obvious descriptions for the same person or organization. When two identities are similar but do not cross a certainty threshold, SNAC relates them to one another and labels them *maybeSameAs*. (SNAC, 2017)

36. Voir par exemple dans le fichier XML associé à Ethel Rosbenberg : https://snaccooperative.org/download/9353978?type=eac-cpf.

Ces précautions[37] expliquent donc pourquoi deux autres entités similaires à Ethel Rosenberg sont reprises sur la page *Similarity Assertions*[38]. Cependant, la liste des non pas deux, mais dix résultats[39] obtenus en recherchant *Ethel Rosenberg* laisse entrevoir l'étendue des efforts devant encore être accomplis afin de réduire la quantité de doublons sur la plateforme[40].

Outre ces défis d'*entity linking* typiques de toute initiative visant à établir des entités uniques à partir de sources éparses, d'autres problèmes de qualité ont été constatés. Certains sont dus à la qualité des données de départ[41], d'autres sont induits par les limites des processus d'extraction des données[42], tandis que d'autres difficultés encore sont liées à la qualité et à la maintenance des liens vers des ressources externes[43]. Par ailleurs, notons encore que la plateforme a dû faire face à des préoccupations d'ordre éthique ou liées à la confidentialité une fois les données mises en ligne[44]. Ces plaintes ont été traitées en respectant les souhaits des individus concernés et en intégrant sur la plateforme un bouton permettant de formuler des demandes de modification (Casalini *et al.*, 2018).

Alors que SNAC continue de proposer de nouvelles fonctionnalités à ses utilisateurs, comme une API (Application Programming Interface) sur laquelle reposent un service de réconciliation[45] et un module d'édition directement reliés au logiciel OpenRefine (Herbert et Hott, 2019), et que des possibilités de téléchargement de données sous forme de RDF sont envisagées (Casalini *et al.*, 2018), l'une des questions qui émerge est celle de la place que la coopérative va accorder à la nouvelle norme de description archivis-

37. Pour d'autres détails sur ce processus de *matching*, voir Pitti *et al.* (2015).

38. https://snaccooperative.org/maybesame/9353978/11342353.

39. https://snaccooperative.org/?count=10&start=0&entity_type=&term=ethel+rosenberg&command=search.

40. Bien que certaines dates de naissance varient légèrement, que des variations puissent être constituées au niveau du nom et que certaines entités semblent concerner le couple Rosenberg dans son ensemble, des fusions pourraient certainement être réalisées entre certaines de ces entités.

41. « [SNAC] is being built using existing description of archival resources. While much of the processing is successful, a significant portion of the data is inaccurate. The resource description data was not created with the current processing objectives in mind. The quality of the resource descriptions is uneven, most often because of simple human error and oversight, either in the data itself or in the encoding of the data. » (SNAC, 2017).

42. « The semantics for batch processing the harvested data was primitive, so only a certain level of precision was possible. Recorded relationships were especially basic, with most defaulting to an *associatedWith* relationship designator » (Casalini *et al.*, 2018).

43. « There are some concerns about maintainng these links, and [we] are working to develop technical solutions for identity verification » (Casalini *et al.*, 2018, p. 28).

44. « These include concernes about identity theft, as well as social issues such as gender identity » (Casalini *et al.*, 2018, p. 28).

45. http://openrefine.snaccooperative.org/.

tique [46] sur laquelle travaille actuellement un groupe d'experts du Conseil International des Archives.

Avant de nous intéresser à cette nouvelle norme en cours d'élaboration, précisons encore que si nous avons choisi de nous pencher en détail sur le cas iconique de la plateforme SNAC, il est clair que ce n'est pas le seul exemple d'implémentation démontrant le potentiel de l'EAC-CPF, comme le documente le site web dédié à ce standard [47].

1.3 Des URIs

1.3.1 Contexte

Comme l'énonce Sibille dans le contexte du développement d'une nouvelle norme internationale de description archivistique, celle-ci ne doit pas seulement répondre à la nécessité qui s'est fait sentir de pouvoir articuler les normes préexistantes [48], mais également à d'autres besoins :

> En effet, il apparaît capital de faire évoluer les descriptions archivistiques traditionnelles pour qu'elles s'adaptent à l'évolution des pratiques de recherche et des technologies du Web sémantique. Disposer d'un modèle unique et global, exploitable informatiquement et adapté aux technologies nouvelles du Web, est un prérequis indispensable pour envisager de faire fructifier des gisements de données archivistiques, alors que celles-ci sont encore beaucoup gérées en silo et que les modes actuels de représentation de l'information restent pauvres. Rapprocher les archives des autres secteurs culturels, dont les travaux de normalisation ont déjà abouti à des modèles internationaux, est un autre enjeu fondamental. (Sibille, 2017, p. 123)

En effet, comme le soulignent Pitti *et al.*, la norme RiC cherche à développer une compréhension plus large du concept de provenance, « by recognizing that records and the people who create, manage, and use them do not exist in isolation but in complex layers of interrelated, interdependent contexts ». C'est à ces enjeux que tente de répondre l'EGAD (Groupe d'experts sur la description archivistique), chargé par l'ICA de développer une nouvelle norme internationale de description archivistique(Pitti *et al.*, 2018).

46. Mise à jour, automne 2020 : le site web dédié à l'ontologie RiC-O fournit des informations supplémentaires à ce sujet, voir : https://ica-egad.github.io/RiC-O/projects-and-tools.html.

47. Voir la page Community and support : https://eac.staatsbibliothek-berlin.de/community-and-mailinglist/.

48.

1.3.2 RiC-CM

Avant de passer au contenu et à la réception des premières versions de travail de cette nouvelle norme, nous proposons une vue synthétique des étapes-clés de son développement jusqu'à aujourd'hui :

- 2008 : le Comité des normes et bonnes pratiques (CBPS) de l'ICA décide d'élaborer un recueil de ses quatre normes de description (Comité des normes et bonnes pratiques, 2019, p. 3)
- 2012 : le CBPS recommande la création d'un modèle conceptuel afin d'éliminer la redondance issue des normes préexistantes (Comité des normes et bonnes pratiques, 2019, p. 5) ; dans la lignée de cette recommandation, l'ICA créé l'EGAD et le charge de développer un modèle conceptuel (Gueguen *et al.*, 2013, p. 571)
- 2013 : l'EGAD décide que deux livrables vont être créés : un modèle conceptuel et une ontologie de haut niveau (Sibille, 2014, p. 6)
- 2016 : l'EGAD publie la version 0.1 de *Records in Contexts Conceptual Model (RiC-CM)* et la soumet aux commentaires (International Council on Archives, 2016c)
- 2019 : l'EGAD publie la version 0.1 de *Records in Contexts Ontology (RiC-O)* et la prévisualisation de la version 0.2 de RiC-CM (International Council on Archives, 2019)
- 2020 : l'EGAD annonce vouloir publier les versions 1.0 de RiC-CM et de RiC-O en novembre, lors du Congrès ICA Abu Dhabi 2020 (International Council on Archives, 2019).

Comme le révèle cette chronologie, la norme RiC repose donc à la fois sur un modèle conceptuel (RiC-CM) et sur une ontologie de haut niveau (RiC-O). Nous nous intéressons d'abord à ce modèle conceptuel, tandis que nous abordons l'ontologie RiC-O au cours de la sous-section suivante.

Ce modèle conceptuel abstrait, s'inspirant de modèles préexistants [49] vise à « considérer les archives, leur histoire et les multiples couches de contextes dans lesquelles elles s'inscrivent » et à « mieux représenter cette complexité » (Clavaud, 2020a, p. 6). Comme le montre la figure 1.3, l'entité *Agent*, qui englobe les personnes, est l'une des quatre principales entités au cœur de ce modèle conceptuel, avec les entités *Record Resource, Instantiation, Activity*. Au total, le modèle RiC-CM repose sur 22 entités qui possèdent des propriétés spécifiques [50] et sont reliées entre elles par 78 types de relations (International Council on Archives Expert Group on Archival Description, 2019b).

Si une présentation et une analyse détaillées de ce modèle dépasseraient le périmètre de notre recherche, notons toutefois que l'entité *Agent* a la parti-

49. Voir Pitti *et al.* (2018).
50. Plus précisément 41 types d'attributs

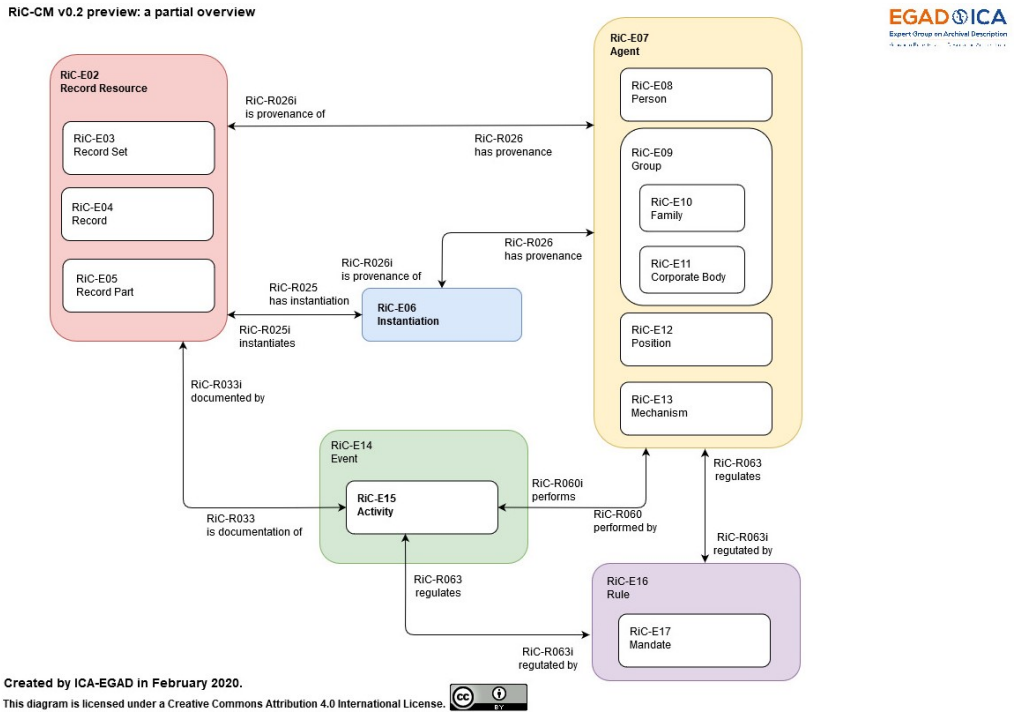


FIGURE 1.3 – Principales entités du Modèle Conceptuel Records in Contexts 0.2. Source : Clavaud (2020a, p. 17) (CC-BY).

cularité d'englober non seulement les personnes et les groupes (qui incluent les familles et collectivités) que l'on retrouve dans la norme ISAAR (CPF), mais également les entités *Position* – c'est-à-dire « the functional role of a Person within a Group » – et *Mechanism* – à savoir, « a process or system created by a Person or Group that performs an Activity » – (International Council on Archives Expert Group on Archival Description, 2019b, p. 11).

Bien que cette norme soit toujours en cours de développement, elle a toutefois déjà fait l'objet de réactions de la part de la communauté archivistique internationale, notamment dans le cadre d'un appel à commentaires lancé par l'ICA en septembre 2016 (International Council on Archives, 2016a). Une soixantaine de parties prenantes – telles que des organisations professionnelles, des laboratoires de recherche, des développeurs de logiciels de gestion d'archives ou encore des particuliers – s'est ainsi exprimée sur la version 0.1 de ce modèle conceptuel *Records in Contexts*. Une vue synthétique de ces plus de 1 000 commentaires issus a été présentée dans le cadre d'une rencontre annuelle de l'EGAD (Timms, 2017).

Nous nous contenterons ici de relever quelques grandes tendances avant de dégager les points les plus saillants relatifs à l'entité *Agent*. Globalement, les commentaires se répartissent entre des considérations générales

et des remarques plus spécifiques portant sur les caractéristiques des entités, des propriétés ou des relations contenues dans la version de travail du modèle conceptuel RiC-CM. Les remarques générales portent tant sur les fondements théoriques du modèle que sur sa portée et son audience, ses principes sous-jacents, sa présentation et sa mise à disposition, son processus de développement, mais aussi la façon dont il est modélisé ou encore la place accordée aux utilisateurs et à leurs besoins, à l'interopérabilité, à l'implémentation réelle et aux *digital records* (Timms, 2017). Les commentaires concernant le contenu du modèle portent principalement sur la façon dont les entités, propriétés ou relations ont été choisies et modélisées, sur leur définition, sur leur portée, ainsi que sur certaines zones de flou ou de chevauchement entre différents concepts.

Une quarantaine de ces commentaires se rapportent à l'entité *Agent* et aux autres entités qui y sont liées. Ils contiennent à la fois des remarques sur les définitions des entités et des suggestions concernant leur modélisation. Ainsi, l'une des demandes visait à ce que soient modélisées de nouvelles entités pour préciser les types d'agents [51], ce qui est désormais pris en compte dans la version RiC-CM 0.2, comme nous l'avons vu au cours des paragraphes précédents. D'autres remarques portent sur le risque de confusion entre *Occupation* et *Position* et suggèrent que l'*Occupation* soit modélisée sous forme de propriété plutôt que d'entité, ce qui est également pris en charge dans le cadre de Ric-CM 0.2 (International Council on Archives Expert Group on Archival Description, 2019b). Enfin, des réserves sont énoncées par rapport aux propriétés liées aux personnes, en particulier les propriétés relatives à l'identité et au genre (Timms, 2017) [52].

Bien que certaines critiques se soient avérées particulièrment vives ou acerbes, l'EGAD se félicite du fait que la communauté archivistique internationale s'intéresse au développement du modèle et participe à son amélioration en fournissant des commentaires détaillés (Timms, 2017). Ainsi, bien qu'il n'ait pas encore communiqué explicitement à ce sujet, le Groupe d'experts a commencé à prendre en compte ces commentaires (Clavaud, 2019a). En témoignent par exemple les réponses apportées aux critiques qui portaient sur le manque de transparence du processus d'élaboration de cette norme, comme la mise à disposition de versions successives de l'ontologie RiC-O, qui sera abordée au cours des prochains paragraphes, ou encore la création d'un dépôt GitHub [53] offrant un libre accès aux fichiers sources de

51. *Person, group, corporate body*.

52. Ce paragraphe est également basé sur la consultation d'un document de travail de l'EGAD auquel nous avons eu accès par l'intermédiaire de l'une de ses membres, Florence Clavaud.

53. Voir : https://github.com/ICA-EGAD/RiC-O.

l'ontologie et de nouvelles opportunités de contribution aux personnes intéressées.

1.3.3 RiC-O

RiC-O est une ontologie OWL pour la description des documents d'archives. Elle constitue une représentation formelle du modèle conceptuel RiC-CM (on Archives Expert Group on Archival Description, 2020). Reposant sur les principes de « usefulness, flexbility, functionality, and extensibility » (Pitti *et al.*, 2018), elle a pour principal objectif de « définir le vocabulaire et les règles applicables aux métadonnées archivistiques ayant la forme de jeux de données RDF » (Clavaud, 2020a).

Il faut noter que l'ontologie RiC-O n'est pas une transposition formelle du modèle RiC-CM. En effet, elle se distingue du modèle RiC-CM dans la mesure où elle contient plus de composants, de manière à pouvoir par exemple affiner la description de certaines entités, les organiser en système polyhiérarchique ou encore leur assigner des attributs supplémentaires (Clavaud, 2020a). Concrètement, la version 0.1 de RiC-O contient près d'une centaine de classes, parmi lesquelles la classe Rico Person [54] – sous-classe de Rico Agent [55] –, qui permet l'encodage de données relatives à des individus. Les instances de cette classe *Person* peuvent ensuite être décrites à l'aide d'une quarantaine de propriétés (outre les propriétés héritées de la classe *Agent*), telles que rico :hasParent, rico :hasBirthDate ou encore rico :isMemberOf, que nous réévoquerons dans le cadre de l'étude de cas présentée en seconde partie.

Bien qu'encore en cours développement, l'ontologie a déjà fait l'objet d'implémentations. C'est en France qu'a été déployée la première mise en œuvre concrète de l'ontologie RiC-O à partir de jeux de métadonnées réels, dans le cadre d'un prototype de visualisation de métadonnées archivistiques (Clavaud, Florence, 2018). L'outil de démonstration PIAAF [56], pour *Pilote d'Interopérabilité pour les Autorités Archivistiques Françaises*, a été développé [57] en France entre 2016 et 2018, dans le cadre d'un projet conjoint – lancé en 2015 – de trois partenaires institutionnels : les Archives nationales de France, le Service interministériel des Archives de France et la Bibliothèque nationale de France (Equipe projet PIAAF, 2018a). Ce prototype applicatif consiste en une interface de recherche et d'exploration dynamique des jeux de données RDF issues de métadonnées archivistiques produites par plusieurs institutions (Equipe projet PIAAF, 2018a). Cette opération, qui fut une

54. https://www.ica.org/standards/RiC/ontology#Person.
55. https://www.ica.org/standards/RiC/ontology#Agent.
56. Voir : https://piaaf.demo.logilab.fr/.
57. Par la société Logilab.

FIGURE 1.4 – Interface PIAAF de visualisation et d'exploration de métadonnées archivistiques converties en RDF. Source : Clavaud et Charbonnier (2020, p. 20) (CC-BY).

première dans le domaine des archives (Clavaud et Château-Dutier, 2017), visait à démontrer :

- la faisabilité de la conversion de ces métadonnées en graphe sémantique (en RDF) [58], conformément à une version très peu avancée de RiC-O
- l'intérêt de l'opération pour la mise en relation de ces différents jeux de métadonnées
- l'intérêt de l'opération pour la recherche dans ce graphe, sa visualisation et la consultation des jeux de données (Clavaud et Charbonnier, 2020, p. 10)

L'équipe à l'initiative de ce projet explique avoir opté pour un modèle de représentation ayant vocation à devenir un standard international – RiC-CM et sa transposition en ontologie RiC-O – plutôt qu'un modèle de référence spécifique au projet, afin de pouvoir prouver que l'opération de conversion était transposable à d'autres jeux de métadonnées (Equipe projet PIAAF, 2018b). Concrètement, ce sont 276 notices d'autorité EAC-CPF et 38 inventaires EAD qui ont été convertis en RDF (Equipe projet PIAAF, 2018b) en utilisant une quinzaine de classes RiC-O [59].

58. « En restituant de façon adéquate la complexité inhérente aux archives et les multiples facettes du contexte archivistique » (Angjeli *et al.*, 2017, p. 158).

59. À savoir : RiC :Agent ; RiC :CorporateBody ; RiC :Person ; RiC :GroupType ; RiC :FunctionAbstract ; RiC :Position ; RiC :Mandate ; RiC :Event ; RiC :AgentName ; RiC :Place ; RiC :Relation ; RiC :Description ; RiC :Record ; RiC :RecordSet ; RiC :Proxy ; RiC :FindingAid (Clavaud et Charbonnier, 2020, p. 13).

La figure 1.4 laisse entrevoir comment le prototype PIAAF affiche ces métadonnées après leur conversion en RDF, avec, sur la gauche, la page [60] dédiée à une entité Personne – à savoir, l'historien et politicien français Jean-Noël Jeanneney – et, sur la droite, la page [61] montrant les archives de cette personne – à savoir ses archives lorsqu'il était directeur de Radio France.

Outre les personnes physiques, le projet a également intégré des notices d'autorité relatives à des collectivités, une démarche qui a mis en lumière les difficultés et défis liés à la représentation du temps. D'une part, des divergences de pratiques ont pu être observées entre les trois institutions [62] ; d'autre part, la combinaison d'un graphe et de la représentation de la dimension temporelle est une pratique encore rare, qui s'avère d'autant plus complexe dans un contexte archivistique, étant donné le nombre important de ressources [63] et de relations devant être croisées (Angjeli *et al.*, 2017, p. 169).

Les résultats du projet, jugés très convaincants, ont permis aux organismes concernés de « mieux comprendre le saut réalisé en termes de précision et de possibilités d'interrogation » (Clavaud, 2020b). En effet, l'approche adoptée visait *une expressivité maximale* (Equipe projet PIAAF, 2018b) ce qui s'est traduit par la volonté de tirer le plus d'informations possibles des jeux de données, y compris lorsqu'elles étaient présentes sous forme implicite [64] (Clavaud et Charbonnier, 2020). Le projet a ainsi contribué à mettre en exergue l'importance de l'indexation :

> Le nombre de ressources RDF *contextuelles* [...] pourrait bien évidemment augmenter fortement si les fichiers de métadonnées source comportent une indexation plus riche et sont plus fortement structurés. Autrement dit, il est important d'indexer plus, et de façon contrôlée, les instruments de recherche et les notices d'autorité. (Equipe projet PIAAF, 2018b)

60. http://piaaf.demo.logilab.fr/resource/FRAN_person_050789.
61. http://piaaf.demo.logilab.fr/resource/FRAN_record-set_050629-top.
62. Ainsi, Angjeli *et al.* ont relevé que « pour la même collectivité, les Archives nationales identifient souvent plus d'entités successives que la BnF ou le SIAF, étant plus sensibles aux changements de fonction, de statut, de position et de nom » (Angjeli *et al.*, 2017, p. 169).
63. À la différence par exemple de l'interface de visualisation de réseaux généalogiques développée dans le cadre de Kindred Britain (https://kindred.stanford.edu/), qui se concentre sur des relations familiales et n'inclue pas de ressources archivistiques, comme le relèvent Clavaud et Château-Dutier (2017).
64. Il s'agissait par exemple de faire émerger les objets ou entités décrits dans les notices d'autorité, comme des noms, événements, lieux ou encore postes occupés. Ainsi, les résultats après la conversion en RDF comptent ainsi plus de 900 agents, alors que les jeux de données initiaux comptaient moins de 300 notices EAC-CPF (Clavaud et Charbonnier, 2020, p. 13 ;p. 16), laissant deviner qu'un grand nombre de ces agents est issu de mentions plus ou moins explicites.

La réalisation du prototype a également montré l'intérêt d'une catégorisation fine des relations (Francart et Charbonnier, 2020), mais aussi les limites du schéma EAC-CPF pour le faire (Equipe projet PIAAF, 2018b). De plus, « le démonstrateur révél[ant] impitoyablement des erreurs subsistantes ou des doublons », il a montré ou rappelé l'importance de la gestion de la qualité des fichiers source (Equipe projet PIAAF, 2018b). Enfin, l'expériméntation a fait apparaître la nécessité pour les institutions patrimoniales de se doter d'identifiants pérennes si elles souhaitent passer au RDF : « des identifiants pérennes pour toutes les ressources RDF à individualiser doivent impérativement être produits en amont de l'opération de conversion, au sein des systèmes d'information » (Equipe projet PIAAF, 2018b).

Dans la continuité du prototype PIAAF, les Archives Nationales de France (ANF) ont souhaité se doter d'un outil permettant la conversion industrielle des métadonnées archivistiques – encodées en EAD et EAC-CPF – en jeux de données RDF conformes à RiC-O. Leur objectif était en effet de pouvoir convertir aisément quelques 15 000 notices d'autorité et 280 000 inventaires d'archives ; c'est chose faite avec RiC-O Converter, un programme reposant sur XLST – un langage informatique de transformation XML – qui permet de lancer cette commande sans compétence particulière et d'obtenir en moins d'une heure les jeux de données RDF correspondants (Francart et Charbonnier, 2020). Développé en 2019-2020 [65], ce logiciel contient des règles spécifiques à l'utilisation des balises EAC-CPF et EAD des ANF et ne couvre donc pas tous les éléments de façon systématique, mais son code [66] est toutefois librement adaptable [67].

Si ces expérimentations ont permis aux ANF de faire émerger des fichiers de départ des entités conceptuelles implicites [68] et de voir que « RiC-O est utilisable pour exprimer la complexité de la description archivistique », elles ont également montré la nécessité d'améliorer et de contrôler la qualité des métadonnées à convertir (Clavaud et Charbonnier, 2020, p. 17).

L'un des principaux points d'attention est ainsi le fait que certaines entités ne possèdent actuellement pas d'identifiant unique. Une solution provi-

65. Dans le cadre d'un marché avec la société Sparna (Francart et Charbonnier, 2020).

66. Code source, exemples et documentation sont librement mis à disposition, voir : https://github.com/ArchivesNationalesFR/rico-converter.

67. Clavaud explique ainsi : « nous avons construit cet outil pour les Archives nationales, mais également dans l'idée qu'il pourrait être utile à tout service d'archives ou tout autre organisme disposant d'inventaires au format EAD et/ou de notices au format EAC-CPF, et souhaitant utiliser les technologies du web de données tout en veillant à la conformité de ses métadonnées avec le standard RiC. C'est pourquoi le logiciel est un logiciel libre. Tout le monde pourra donc s'en servir et en modifier le code pour l'adapter à ses besoins. » (Clavaud, 2020b).

68. C'est le cas par exemple de certains types de relations dans le cadre des fichiers EAC-CPF, voir (voir Francart et Charbonnier, 2020, p. 17).

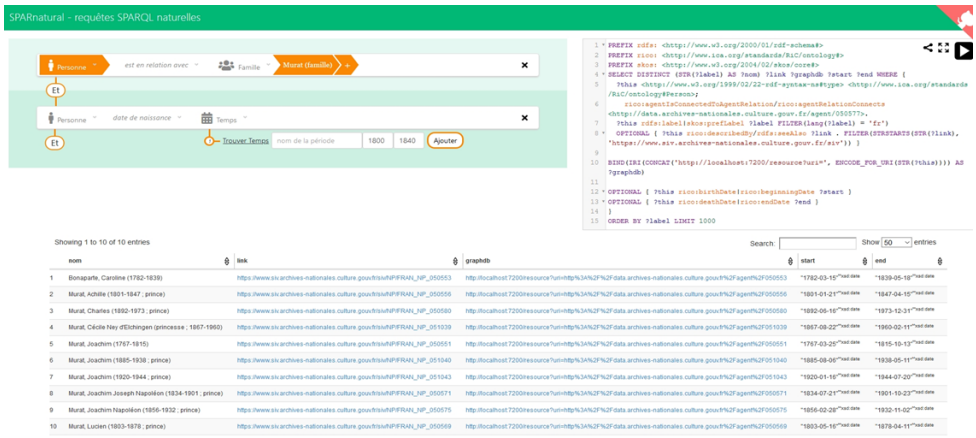


FIGURE 1.5 – Interface d'exploration du graphe RDF des producteurs d'archives des Archives nationales de France. Source : Francart et Charbonnier (2020, p.44) (CC-BY).

soire [69] a été déployée. Si, dans le cadre d'entités appartenant à la classe RiC-O Agent [70] les URIs – fictifs et provisoires – s'avèrent relativement concis [71], ils s'annoncent par exemple plus complexes dans le cadre de la classe RiC-O Agent Name [72]. En effet, ils sont dans ce cas destinés à être composés d'une concaténation de l'identifiant de la notice d'autorité, de l'encodage de la forme autorisée du nom et de ses dates d'utilisation [73] (Francart et Charbonnier, 2020, p. 21), laissant entrevoir les défis de création et de maintenance qu'entraîne une utilisation systématique d'URIs, notamment dans le cas de modifications des données...

Enfin, comme le souligne Clavaud (2020b), « disposer de jeux complets de métadonnées archivistiques en RDF, c'est bien. Mais il faut encore mettre en place des outils d'interrogation et de visualisation de ces données ». Dans le cadre du projet RiCO Converter, des expérimentations ont ainsi été réalisées afin que le graphe de données RDF résultant des opérations de conversion puisse être exploré de façon intuitive. La figure 1.5 montre comment une interface graphique conviviale – basée sur Sparnatural [74] – permet de sélectionner des valeurs au sein de listes déroulantes, qui vont se charger de générer les requêtes SPARQL correspondantes.

69. Étant donné que « les Archives nationales n'ont pas encore défini de stratégie pour la construction et la gestion des URIs » (Francart et Charbonnier, 2020, p. 19).
70. Voir *rico :Agent* : https://www.ica.org/standards/RiC/ontology#Agent.
71. Ils reposent sur l'identifiant unique qui avait été attribué à la notice d'autorité, résultant par exemple en : http://data.archives-nationales.culture.gouv.fr/agent/51126.
72. Voir *rico :AgentName* : https://www.ica.org/standards/RiC/ontology#AgentName.
73. Par exemple : http://data.archives-nationales.culture.gouv.fr//agentName/000005-Minist%C3%A8re%20de%20la%20Culture-19950518-19970602.
74. Sparnatural est un composant d'exploration open-source basé sur SPARQL (Francart et Charbonnier, 2020), voir : https://github.com/sparna-git/Sparnatural/.

Enfin, notons qu'au-delà de ces premières expérimentations menées par les Archives nationales de France [75], de nouvelles initiatives sont actuellement en cours de développement ailleurs dans le monde [76]. De plus, il est certain que les services d'archives n'ont toutefois pas attendu que le développement de cette nouvelle norme arrive à son terme pour se lancer dans des expérimentations liées à la transformation de descriptions archivistiques en *Linked (Open) Data*. Bien que ces initiatives soient encore peu nombreuses en regard des multiples projets qui ont essaimé dans le secteur des bibliothèques au cours des dernières années [77], Niu (2016) a toutefois dressé en 2016 un panorama de 17 projets ayant mis en œuvre des données liées archivistiques.

Nous nous concentrons ici sur trois projets visant à interconnecter des ressources autour d'un individu : Linking Lives, The Open Memory Project et War Sampo.

En 2011, le projet Linking Lives (UK) a décidé de tirer parti des descriptions archivistiques ayant été publiées sous forme de Linked Open Data dans le cadre du projet Linked Open Copac Archives Hub (LOCAH) [78] et de lancer une nouvelle expérimentation. Le projet Linking Lives avait pour objectif principal d'explorer de nouvelles façons de présenter les données liées au profit des chercheurs, en se centrant autour des personnes :

> We should recognise that researchers may not just be interested in archives. Indeed, they may not really have thought about using primary source material, but they may be very interested in biographical information, known and unknown connections, events during a person's lifetime, etc. We want to show that archives can benefit from being presented not in isolation, but as a part of all of the diverse data sources that can be found to create a full biographical picture, and to enable researchers to make connections between people and events to create different narratives. (Linking Lives, 2011a)

75. Pour d'autres exemples encore et une vision plus large des efforts en cours, voir Clavaud (2019b).

76. Pour une liste récapitulative et régulièrement mise à jour de ces projets, voir la page *RiC-O projects and tools* sur le site web dédié à l'ontologie RiC-O : https://ica-egad.github.io/RiC-O/projects-and-tools.html.

77. Comme https://data.bnf.fr/ en France ; https://data.bibliotheken.nl/ aux Pays-Bas ; https://www.dnb.de/EN/lds en Allemagne ; le projet LiDa au Luxembourg https://data.bnl.lu/apis/lida/, pour ne citer que quelques exemples issus des bibliothèques nationales des pays frontaliers de la Belgique.

78. Diverses considérations liées à la conversion de données EAD vers du RDF peuvent être consultées sur la version archivée du blog associé au projet, (voir par exemple Johnston, 2011; Stevenson, 2011).

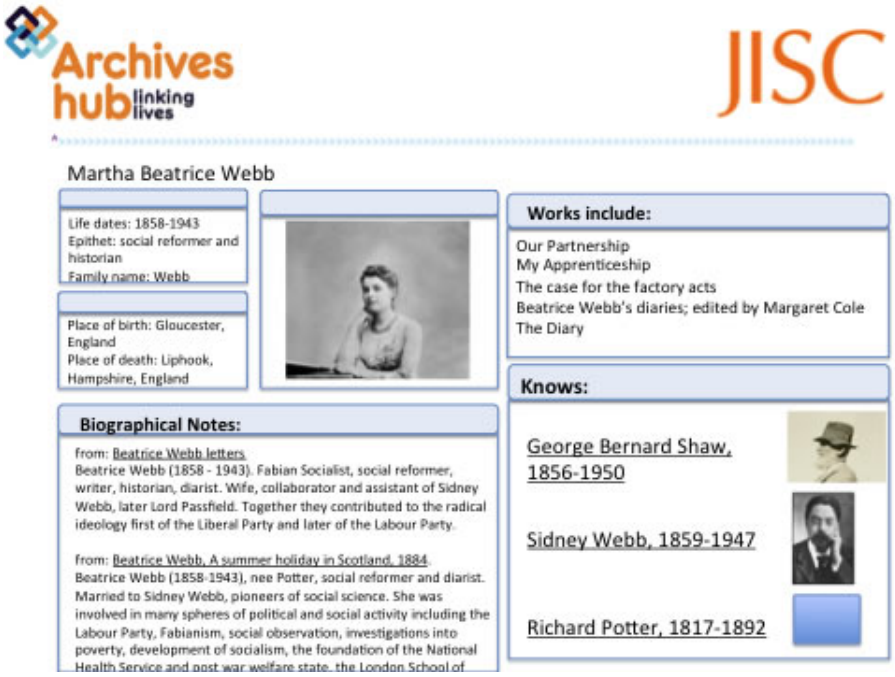


FIGURE 1.6 – Interface du projet Linking Lives. Source : Stevenson (2012b) (CC BY-NC).

La figure 1.6 montre un aperçu de la façon dont l'interface développée dans le cadre de ce projet devait permettre de publier une page par individu, de le doter d'un identifiant unique, d'afficher divers types d'informations biographiques et d'ensuite enrichir cette page à l'aide de données externes liées, que ce soit sous forme de textes, d'images ou de liens, en rendant explicite la source des données (Stevenson, 2012a).

Stevenson a mis en exergue les enjeux liés aux notions de provenance, de confiance et d'authenticité dans le cadre de tels projets (Linking Lives, 2011a) et a également souligné la nécessité de pouvoir fournir davantage de preuves des bénéfices que de tels dispositifs peuvent offrir à des utilisateurs finaux qui ne connaitraient ni SPARQL ni RDF (Stevenson, 2012a). À l'issue du projet, elle relève trois principales difficultés : le manque d'outils et de modèles propres à un domaine émergent ; les difficultés liées aux données de départ [79] ; les défis liés à l'utilisation de sources de données externes

79. Voir Linking Lives (2011b). Dans le même esprit, nous pouvons également penser au projet Traces Through Time, mené par the National Archives (UK) en 2014-2015, qui a mis en lumière les difficultés liées à l'automatisation de l'extraction de noms de personnes de documents d'archives semi ou non structurés, et de leur interconnexion (voir Ranade, 2015, 2016b) ; ou encore au projet de dédoublonnage des notices de personnes du Petőfi Museum of Litterature (Bánki *et al.*, 2016).

(Stevenson, 2012a). Ce dernier point renvoie à la question de la maintenance :

> The linked data space is constantly changing, so new data is created all the time and improvements are made to existing data. This makes it quite a moveable feast, and you have to make decisions about whether to go back to updated datasets and re-examine them, or stick with what you have. This may be a challenge in terms of maintaining the interface. We may find that the need to monitor the linked data space takes up significant time. We will continue to maintain our linked data interface, and seek to add some more external data sources, and then we will monitor the result, see how much it is used, and how much effort we have to invest in ensuring it is current and all links are operable. (Stevenson, 2012a)

Ces derniers mots ont aujourd'hui une résonnance particulière, sachant que la plateforme de publication des données n'a plus été mise à jour depuis 2013 et est aujourd'hui uniquement consultable dans une version archivée [80]. Cependant, un nouveau projet traitant des noms de personnes a été lancé par The Archives Hub au printemps 2020, témoignant d'un intérêt renouvelé pour cette problématique [81].

The Open Memory Project est un autre projet au carrefour entre personnes, archives et Linked Open Data. Réalisé par la Fondation CDEC (Centro di Documentazione Ebraica Contemporanea) à Milan – en partenariat avec la compagnie Regesta.exe –, ce projet a remporté le challenge LODLAM [82] 2015 (Regesta.com, 2015). Son point de départ est une base de données, The Names of the Victims of the Shoah in Italy, qui est composée de plusieurs milliers de noms [83] et d'une trentaine de champs contenant des informations biographiques, mais également des données sur la persécution des victimes et de leurs proches (Brazzo et Mazzini, 2015). Ces informations ont été structurées autour d'entités Personne à l'aide de vocabulaires préexistants – comme FOAF ou BIO – ainsi qu'à l'aide d'une ontologie créée spécifiquement pour décrire les concepts et relations caractérisant les activités liées à la persécution des Juifs en Italie entre 1943 et 1945. Comme le

80. Voir : http://data.archiveshub.ac.uk/.

81. Voir les posts de blogs dédiés au Names Project : https://blog.archiveshub.jisc.ac.uk/tag/names-project/.

82. LODLAM pour Linked Open Data in Libraries, Archives, and Museums, voir : https://lodlam.net/challenge/.

83. Plus de 19 000 (Brazzo et Mazzini, 2015), mais seules 9 009 victimes de la Shoah sont déjà présentes sur http://digital-library.cdec.it/cdec-web/persone au mois de septembre 2020.

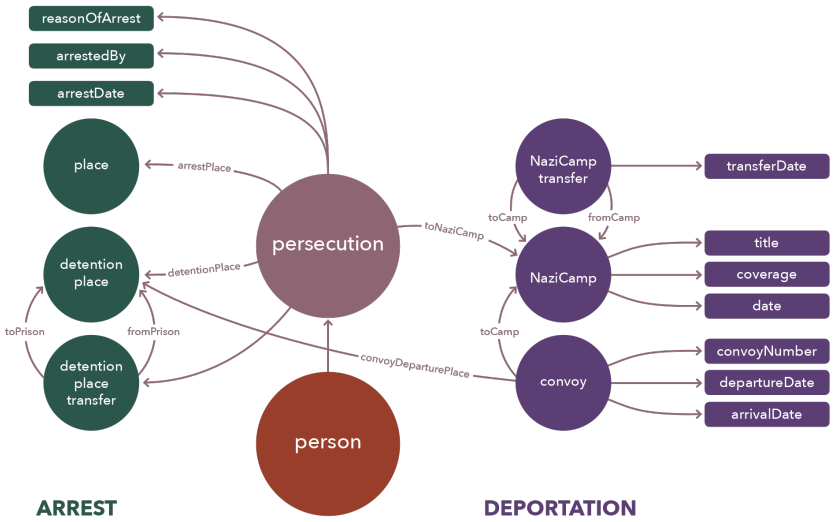


FIGURE 1.7 – The Shoah domain ontology. Source : CDEC et Regesta.exe (CC-BY) (http://dati.cdec.it/lod/shoah/reference-document.html).

montre la figure 1.7, l'*épine dorsale* de cette ontologie est la relation entre les classes *Persécution* et *Personne* (dati.CDEC, 2015).

En parallèle de la modélisation de ces informations, chacune des entités Personne a ensuite été connectée aux diverses ressources du CDEC – archives historiques, archives photographiques, archives audio-visuelles, collections de bibliothèque, données de recherche historique – ainsi qu'à des entités externes issues du Web de données – comme DBpedia – à l'aide de la relation d'équivalence *owl :sameAs* (Brazzo et Mazzini, 2015). À noter que les identifiants uniques associés à chacune des entités jouent un rôle crucial dans le cadre de la connexion de chaque entité aux archives de l'institution : Brazzo et Mazzini expliquent qu'ils sont venus *remplacer* les traditionnels fichiers d'autorité EAC-CPF attendus dans le cadre de descriptions archivistiques (Brazzo et Mazzini, 2015).

À l'instar du projet Linking Lives, l'Open Memory Project permet donc aux utilisateurs effectuant des recherches sur des personnes d'avoir sous leurs yeux des données biographiques, des informations sur des relations familiales, ainsi que des renvois vers des ressources tant internes – comme l'illustre la figure 1.8 montrant l'entité Primo Levi – que externes. Les utilisateurs peuvent consulter les données dans divers formats [84] et disposent

84. Voir : http://dati.cdec.it/indiceEN.html.

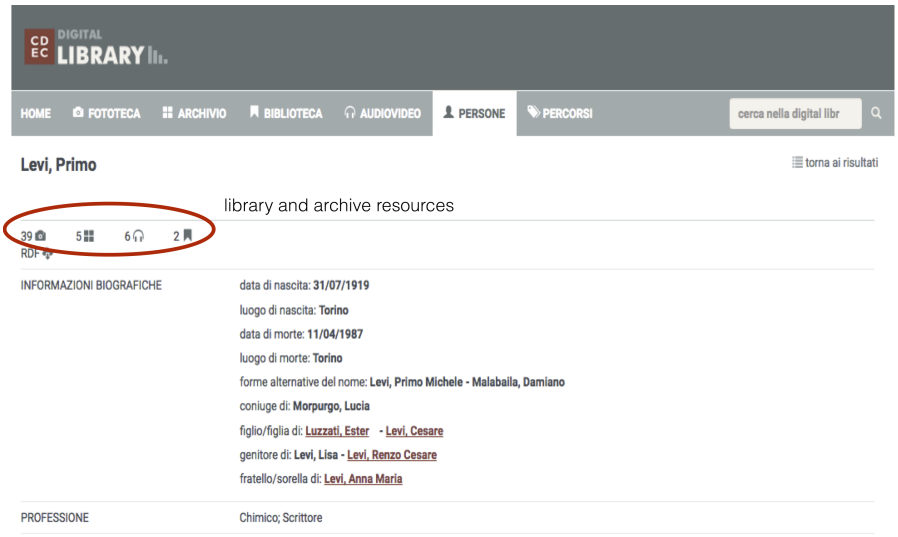


FIGURE 1.8 – Primo Levi et les ressources qui lui sont associées sur le catalogue en ligne du CDEC. Source : CDEC (http://digital-library.cdec.it/cdec-web/persone/detail/person-5002/levi-primo.html).

également d'un point d'accès SPARQL [85] leur permettant d'effectuer des requêtes plus complexes visant par exemple à récupérer la liste de tous les Juifs d'Italie ayant été déportés en tant qu'opposants politiques [86].

Également dans le domaine de la Seconde Guerre mondiale, mais cette fois-ci à travers une perspective militaire, le projet finlandais WarSampo, lancé en 2014, cherche à harmoniser et mettre à disposition [87] de larges collections de données et ressources liées à la guerre sous forme de Linked Open Data. Ce projet, qui a gagné le challenge LODLAM 2015 (Aalto University, 2017), s'inscrit dans le cadre plus large de l'initiative Linked Data Finland, qui est menée par le Semantic Computing Research Group de l'Université Aalto en collaboration avec l'Université d'Helsinki et un consortium d'une vingtaine de partenaires issus du service public et privé (Linked Data Finland, 2020). Concrètement, la base de connaissance WarSampo est composée d'un service de données liées [88] et d'une plateforme sémantique [89] qui propose un accès à partir de différentes perspectives, comme des événements, des places ou encore des personnes liées à la guerre. Cette dernière perspective [90], centrée autour d'environ 100 000 individus (Koho *et al.*,

85. http://lod.xdams.org/sparql.
86. Voir exemples (en italien) : http://dati.cdec.it/esempi.html.
87. Des chercheurs, mais également du grand public (Koho *et al.*, 2019b).
88. http://www.ldf.fi/dataset/warsa.
89. http://sotasampo.fi/.
90. https://www.sotasampo.fi/en/persons/.

FIGURE 1.9 – Exemple d'entité Personne sur la plateforme WarSampo. Source : WarSampo (https://www.sotasampo.fi/en/persons/person_50).

2019a), permet un accès centralisé à divers types de données et ressources – comme le montre l'aperçu proposé en figure 1.9.

WarSampo combine des données issues d'une douzaine de sources hétérogènes – des images, des fichiers tabulaires et des documents PDF, contenant des données plus ou moins détaillées –, provenant notamment des Archives nationales de Finlande [91]. Pour agréger ces informations, un important réseau de relations – comptabilisant plus de 14 millions de triples (Koho *et al.*, 2019b) – a dû voir le jour, comme le montre – partiellement [92] – la figure 1.10. Cette étape d'agrégation est possible grâce à l'harmonisation préalable des données préexistantes selon une ontologie personnalisée reposant principalement sur le modèle CIDOC CRM (Conceptual Reference Model) [93]. Cela signifie qu'à quelques exceptions près [94], toutes les données sont modélisées [95] autour de la notion d'événement [96].

91. Ainsi, la base de données des Archives nationales concernant les victimes de guerre contient à elle seule des données sur près de 95 000 personnes (Koho *et al.*, 2019b).

92. Les informations temporelles ne sont par exemple pas reprises.

93. http://www.cidoc-crm.org/.

94. Comme l'expliquent Hyvönen *et al.* : « Person instances record only the basic properties, like family name (the only required property), forenames, a description, and provenance data, i.e., a link to the source from which the data was extracted. All other information is modeled as events. » (Hyvönen *et al.*, 2016).

95. Pour une description détaillée de cette étape, voir Leskinen *et al.* (2017); Koho *et al.* (2019b).

96. Koho *et al.* en expliquent ainsi les raisons, mais aussi les limites : « Event-based modeling is an effective approach to representing wars, enabling the harmonization of heterogeneous data, that can be used in spatio-temporal analytics and user interfaces without the need to adjust the queries for each source dataset separately. The downside of using an

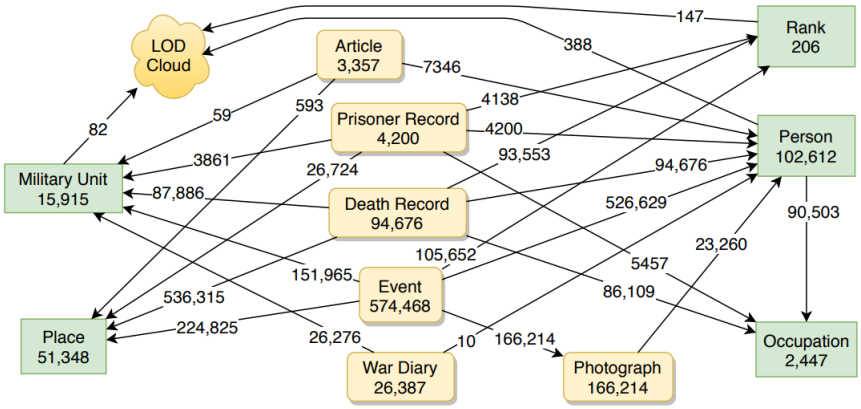


FIGURE 1.10 – Vue d'ensemble des principaux types d'entités sur la plateforme WarSampo et de leurs relations à d'autres entités (en vert) ou à des jeux de métadonnées en RDF (en jaune). Source : Koho *et al.* (2019b, p.6).

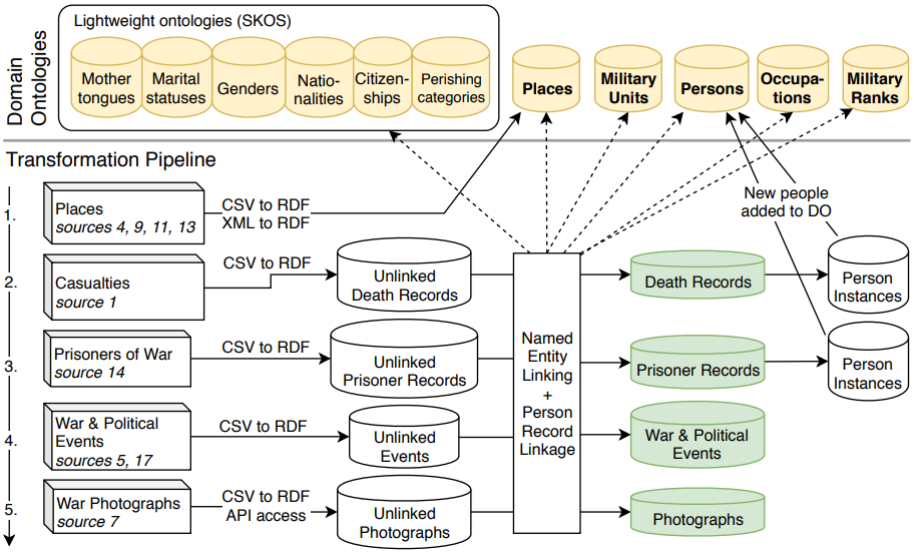


FIGURE 1.11 – Processus de transformation des données WarSampo. Source : Koho *et al.* (2019b, p.9).

Comme le soulignent Koho *et al.*, l'un des défis dans le cadre de ce type de plateforme c'est la maintenance des données et de leurs interconnexions. Ils expliquent ainsi qu'en raison des difficultés liées à l'édition directe des données modélisées selon le modèle CRM, ils préfèrent assurer la maintenance des jeux de données d'origine, qui font ensuite l'objet de toute une série d'opérations de transformation, de manière automatisée. La figure 1.11 montre un aperçu de ce dispositif, qui permet qu'à chaque mise à jour des fichiers CSV d'origine, les modifications puissent être facilement réintégrées sur WarSampo (Koho *et al.*, 2019a,b). Bien qu'elles puissent être automatisées, ces étapes représentent parfois d'importants défis. Koho *et al.* mettent ainsi en exergue les difficultés que peut représenter le *mappage* de certaines données : « covering and disambiguating all military ranks is clearly a simpler task than performing the same task with all wartime places [97]. In general, it is not realistic to assume that the domain ontologies completely cover their domain » (Koho *et al.*, 2019a, p. 3). Le degré de succès de tels processus dépend également de la richesse des données disponibles [98] et de leur taux de complétude.

Enfin, le caractère flexible de la plateforme WarSampo lui permet d'intégrer de nouveaux jeux de données, comme par exemple ces données sur des prisonniers finlandais de la Seconde Guerre mondiale issues de sources multiples. L'intégration de ces données contenant parfois des valeurs multiples et contradictoires pour une même propriété fut l'occasion d'adapter la plateforme, afin de pouvoir prendre en compte les données conflictuelles tout en spécifiant la source de l'information [99] (Koho *et al.*, 2019a) ; en revanche, aucune stratégie ne semble avoir été développée pour prendre en charge – de façon structurée – l'incertitude [100]. Il faut par ailleurs noter que l'appli-

event-based model for all the datasets is its complexity and verbosity : photographs are, for example, modeled as an image and an event creating it, which can lead to complex and slow queries. » (Koho *et al.*, 2019b).

97. Dans une autre publication, ils expliquent « In entity linking, disambiguating some entity types without much context information has been found difficult. For example, place names are usually unambiguous on the municipality level, but automatically disambiguating the names of villages, farms, and lakes can not be done reliably due to high amount of synonymy. » (Koho *et al.*, 2019b, p. 11).

98. Le taux d'informations disponibles pouvant énormément varier : « Some data sources, like the casualties database, provide detailed descriptions of person's life span, places, profession, marital status, etc. In contrast, sources such as the Organization Cards might only mention that, e.g., someone called Captain Karhunen has been in command of his unit in a certain battle » (Hyvönen *et al.*, 2016, p. 7).

99. Pour stocker cette information, « the approach used with the prisoners of war dataset is storing source information using RDF reification with the DCMI Metadata Terms property source. » (Koho *et al.*, 2019a, p. 3).

100. De façon peut-être anecdotique, nous avons relevé sur une capture d'écran de la plateforme WarSampo (voir la figure 2 de Koho *et al.*, 2019a) qu'une des valeurs associée à la propriété *nombre d'enfants* était *[4 ?]*.

cation [101] créée spécifiquement pour exploiter ce jeu de données offre une possibilité de *recherche à facettes*, ces dernières constituent une voie d'accès alternative au contenu de la base de connaissance, accessible au grand public sans qu'il ne doive pour autant maîtriser l'art des requêtes SPARQL [102]. Si cette application offre un riche aperçu de la manière dont les Linked Open Data peuvent être rendues accessibles au grand public, en revanche elle soulève la question de la polyvalence du dispositif. Dans ce cas-ci, des interfaces distinctes ont en effet été développées pour les données sur les personnes en général, pour les données concernant des prisonniers de guerre ou encore pour les données relatives à des victimes de guerre [103].

Finalement, à travers des exemples comme le prototype PIAAF ou la plateforme WarSampo, nous avons vu la façon dont le secteur des archives se saisit des technologies du Web de données, faisant des personnes des entités centrales autour desquelles sont agrégées des données issues tant de sources internes que externes. Cela nous a permis de constater que ce genre de plateforme accueille généralement une copie de données elles-mêmes encodées et maintenues au cœur de systèmes d'information distincts. Dans le chapitre suivant nous explorerons comment un outil tel que Wikibase permet aux institutions patrimoniales de créer, éditer et publier en un même espace des *Linked Data*.

101. Il s'agit de la perspective *Prisoners of War* : https://www.sotasampo.fi/en/prisoners/.
102. Cette fonctionnalité est implémentée par le biais de l'outil SPARQL Faceter : « an HTML based component tool on the client side that can be plugged on virtually any public SPARQL endpoint on the web, using only SPARQL API for data retrieval » (Koho *et al.*, 2016).
103. Voir les différentes *perspectives* proposées en page d'accueil de la plateforme WarSampo : https://www.sotasampo.fi/en/.

2 | Gestion des données structurées avec Wikibase

Introduction

Ce chapitre s'intéresse à l'état de l'art concernant la gestion des données d'autorité dans le contexte du patrimoine culturel, en mettant l'accent sur le logiciel libre Wikibase – *an open-source software suite for creating collaborative knowledge bases* [1]. Considéré comme l'un des logiciels les plus populaires, puissants et utiles par les experts du Web sémantique (Hitzler, 2020, p. 6), il offre aujourd'hui la possibilité à des institutions ne disposant que de peu de ressources de publier des données multilingues sous forme de Linked Open Data (Fauconnier, 2018), mais également de bénéficier de fonctionnalités facilitant leur maintenance. En effet, comme souligné dans l'introduction, lorsque de telles fonctionnalités sont directement intégrées à un outil, elles rendent possible une meilleure gestion des données (Lovins et Hillmann, 2017). C'est le cas par exemple du *versioning*, qui permet un suivi précis des modifications des données, et ce, à un degré de granularité très fin.

Si Mattern (2018) a mis en exergue le fait que « maintenance at any particular site, or on any particular body or object, requires the maintenance of an entire ecology » : il est intéressant de relever ici que Wikibase, qui peut être envisagé comme l'infrastructure centrale de l'*ecology* nécessaire à la maintenance des données d'autorité, s'inscrit lui-même dans un écosystème plus large. En effet, comme nous le verrons au cours des prochains paragraphes, bien que Wikibase ait d'abord été créé comme un outil destiné à satisfaire les besoins du projet Wikidata, il aspire aujourd'hui à devenir un véritable écosystème [2]. En effet, l'idée n'est pas de créer de façon isolée

1. https://wikiba.se/.

2. Ainsi, lors de l'édition 2020 du FOSDEM (Free and Open Source Software Developers' European Meeting) – une conférence qui a lieu annuellement entre les murs de l'Université libre de Bruxelles –, au cours d'une session dédiée aux *Collaborative Information and Content Management Applications*, Lydia Pintscher (*product manager* pour Wikidata) dédie sa

ses propres données structurées lisibles par des machines, mais de prendre place au sein d'un écosystème préexistant. Ce dernier est composé de Wikidata, dont les données alimentent tant les projets Wikimedia (comme Wikipedia ou Wikimedia Commons) que des applications externes, et d'instances Wikibase pouvant être chacune connectée à Wikidata, mais également entre elles.

Comme l'explique Pintscher, le développement de Wikibase est né du besoin de pouvoir stocker des données spécialisées qui n'avaient pas leur place sur Wikidata [3]. En effet, dans un idéal de Web décentralisé, cette base de connaissance n'est pas appelée à contenir – et surtout maintenir ! – toutes les données connues sur notre monde [4]. Les données d'autorité des archives, musées et bibliothèques relevant a priori aussi des données spécialisées, une relation mutuellement enrichissante devrait pouvoir dès lors être envisagée entre ces institutions, Wikidata et l'écosystème Wikibase (Pintscher *et al.*, 2019b, p. 10). C'est ce que nous aborderons au cours de ce chapitre, qui revient d'abord sur l'origine et l'évolution de ce logiciel, avant d'aborder ses possibilités et ses limites, en faisant appel à des sources aussi variées que des rapports d'activités, des offres d'emploi, des documents stratégiques, des rapports de bugs ou encore des échanges d'utilisateurs prenant place dans le cadre d'applications de messagerie instantanée.

2.1 Écosystème Wikibase

2.1.1 Origines

Cette sous-section revient sur les origines de Wikibase, quand il s'agissait alors de discrètes lignes de code à l'ombre des projecteurs et destinées avant tout à mieux répondre aux besoins de l'encyclopédie en ligne Wikipédia.

Tout cela nous ramène en 2012, année du lancement de la base de connaissance Wikidata. Wikimedia Deutschland – le *chapitre* allemand du mouvement Wikimedia [5]. – et la Wikimedia Foundation préparent le lance-

présentation à Wikibase, qu'elle n'intitule pas simplement *Wikibase*, mais *Wikibase Ecosystem* (Pintscher, 2020).

3. Nous reviendrons au cours des pages suivantes sur les critères établis par Wikidata pour traiter cette question.

4. Lors d'une présentation donnée dans le cadre de la conférence internationale Wikimania 2019, Pintscher et Ohlig attirent l'attention sur la notion informatique de *single point of failure* : si Wikidata avait l'ambition de contenir l'ensemble des données spécialisées de toutes les disciplines, une panne de cette base de connaissance représenterait alors un risque important pour le reste du système qui en dépend – à savoir l'ensemble des systèmes et applications alimentées par Wikidata (Pintscher et Ohlig, 2019, p. 7).

5. Il s'agit de la plus ancienne et de la plus grande des 40 sections indépendantes du mouvement Wikimedia. Elle compte, début 2020, 140 employés et 74 000 membres (Wikimedia Deutschland, 2020).

ment d'une base de connaissance collaborative destinée à documenter *the world's knowledge* et dont l'objectif est de pouvoir soutenir les plus de 280 éditions linguistiques de Wikipédia à l'aide de données structurées :

> Wikidata is expected to lead to a higher consistency and quality within Wikipedia articles, as well as increased availability of information in the smaller language editions. At the same time, Wikidata will decrease the maintenance effort for the 90,000 volunteers editing Wikipedia. (Matthew, 2012)

Quelques jours plus tard, le magazine Atlantic titre ainsi : « Teaching Wikipedia to Write Itself », soulignant bien l'enjeu à l'aide d'un exemple tablant sur l'élection d'un nouveau Président de la République française :

> This isn't a question of politics but of information – how does the world's information sources come to reflect a French changing of the guard ? For Wikipedia, one of the web's largest and most up-to-date compendia of facts, the answer is complicated. It's not just the Wikipedia entry for « France » or « President of France » that requires quick updating, but many, many more pages that reference Sarkozy, and not just in English or French but in the more than 280 languages in which Wikipedia appears. Who will update all those references ? Currently, some 90 000 volunteer Wikipedia editors. Those heroes of the collaborative web do the yeoman's work that keeps Wikipedia updated. But could there be an easier way ? Could all of those changes happen automatically ? (Rosen, 2012)

Il s'agit donc de rassembler au sein d'une même infrastructure des données structurées lisibles tant par des humains que par des machines, de manière à pouvoir mettre à jour de façon automatisée, centralisée et synchronisée certaines parties d'une même page dans ces variantes linguistiques des éditions Wikipédia. Pour y parvenir, le développement initial de Wikidata, réalisé par une équipe de huit développeurs, est organisé en trois phases : la première vise à centraliser les liens entre les différentes éditions linguistiques de l'encylcopédie ; la deuxième consiste à rendre Wikidata éditable et utilisable ; la troisième doit permettre la création automatisée de listes ou graphiques alimentés par les données de Wikidata [6] (Matthew, 2012).

6. C'est notamment en cela que Wikidata se démarque du projet DBpedia (https://wiki.dbpedia.org/). Lancé en 2007, ce dernier vise à extraire les informations – structurées et multilingues – de Wikipédia afin de les rendre accessibles à tous sur le Web en utilisant les technologies du Web sémantique et des données liées (Lehmann *et al.*, 2015). Si DBpedia se concentre sur l'extraction d'information et l'interconnexion de données – elle a été décrite comme l'un des points centraux du *Linked Open Data cloud* (Lehmann *et al.*, 2015) –, Wikidata se propose pour sa part de réinjecter de l'information dans Wikipédia.

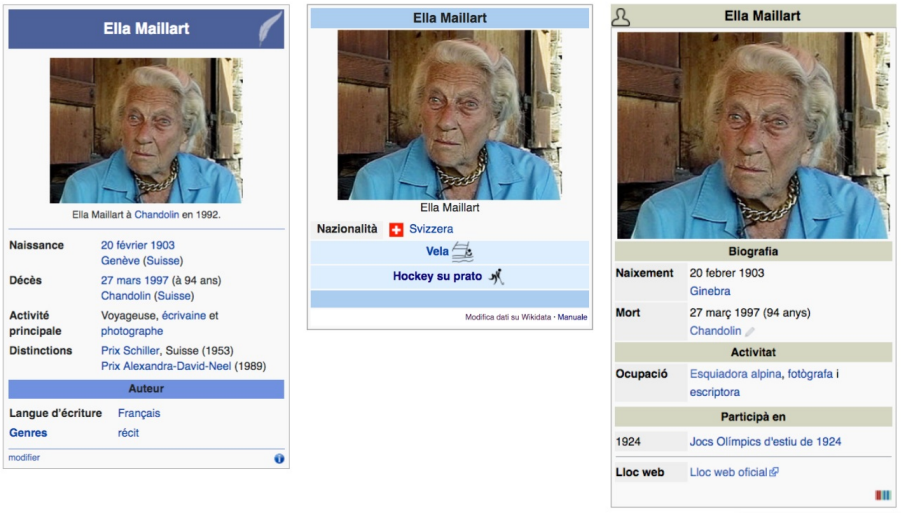


FIGURE 2.1 – Infoboxes dédiées à Ella Maillart issues des articles Wikipédia en français, italien et catalan. Source : Wikipédia (novembre 2019).

Il faut toutefois noter que si sur papier la vision à l'origine du projet Wikidata semble relativement limpide et réalisable – des données centralisées et synchronisées via Wikidata –, dans les faits le résultat est plus nuancé. À savoir que si la première étape visant l'interconnexion des différentes éditions linguistiques est pleinement implémentée depuis mars 2013 [7], en revanche toutes les versions linguistiques ne tirent actuellement pas profit des possibilités d'harmonisation, de centralisation et de synchronisation des informations présentes dans les infoboxes [8] via Wikidata.

Comme le montre la figure 2.1, le contenu des infoboxes dédiées à l'exploratrice suisse Ella Maillart diffère par exemple en fonction de l'édition Wikipédia consultée. Ainsi, sur la première infobox, en français, elle apparaît comme une *voyageuse, écrivaine et photographe* ; sur la deuxième, en italien, nous apprenons qu'elle faisait de la voile et du hockey sur gazon ; tandis que la troisième, en catalan, nous apprend qu'en plus d'être écrivaine et photographe, elle avait également comme occupation le ski alpin. Ces trois exemples illustrent des pratiques qui ne sont pas harmonisées : sur la page Wikipédia en français [9], c'est une *infobox écrivain* [10] qui est utilisée, de façon

7. En témoigne la section *dans d'autres langues* affichée au bas de la barre latérale accompagnant les articles Wikipédia, directement alimentée par Wikidata (et plus précisément par les liens documentés dans les *sitelinks*).

8. Une infobox est une « table de données présentant sommairement des informations importantes sur un sujet [...], placée en général en haut à droite de l'article » (Wikipédia, 2019).

9. https://fr.wikipedia.org/wiki/Ella_Maillart.

10. https://fr.wikipedia.org/wiki/Modèle:Infobox_Écrivain.

totalement indépendante de Wikidata [11] ; sur la page Wikipédia en italien [12], c'est une infobox *Sportivo* [13] qui est utilisée, combinant des informations encodées manuellement et des informations issues de Wikidata (comme par exemple la photo, la taille et le poids de l'athlète) ; enfin, sur la page Wikipédia en catalan [14], c'est une infobox *persona* [15], qui remplace les anciennes infoboxes spécialisées et qui est, par défaut, intégralement alimentée par Wikidata [16].

Cet exemple révèle les disparités dans l'interconnexion entre les différentes variantes linguistiques de Wikipédia et Wikidata. Ainsi, si la communauté derrière Wikipédia en catalan fait partie des pionnières en la matière, ayant créé des modèles d'infoboxes incluant des cartes interactives ou encore des informations biographiques détaillées et ayant annoncé en 2017 que plus de 50% de ses 550 000 articles utilisaient des données provenant de Wikidata (Estermann, 2018), il faut savoir que toutes les communautés ne partagent pas le même enthousiasme. Un phénomène expliqué par de la résistance au changement et de la méfiance vis-à-vis de Wikidata et de sa qualité (Good *et al.*, 2016; Alvarez, 2019) [17].

En parallèle de ce processus de création de liens interwikis et d'intégration de données Wikidata dans des articles Wikipédia, la base de connais-

11. Précisons toutefois que de plus en plus d'articles de Wikipédia en français (voir par exemple l'article dédié à Andrée de Jongh : https://fr.wikipedia.org/wiki/Andrée_De_Jongh) utilisent désormais le canevas *Infobox Biographie 2* (voir : https://fr.wikipedia.org/wiki/Modèle:Infobox_Biographie2), qui est directement alimenté par Wikidata.

12. https://it.wikipedia.org/wiki/Ella_Maillart.

13. https://it.wikipedia.org/wiki/Template:Sportivo/man.

14. https://ca.wikipedia.org/wiki/Ella_Maillart.

15. https://ca.wikipedia.org/wiki/Plantilla:Infotaula_persona/ús.

16. Il est cependant toujours possible de configurer ce qui provient par défaut de Wikidata et ce qui est ajouté manuellement (Alvarez, 2019).

17. C'est le cas par exemple de certains membres de la communauté Wikipédia francophone ayant eu l'occasion de se prononcer sur l'utilisation des données Wikidata dans des articles Wikipédia en français et s'exprimant sans détour, à l'instar de l'utilisateur *Zapotek* : « À ma connaissance, rien (aucune expérience préalable) ne permet de penser que des contributeurs de projets différents et pouvant parler des langues différentes sauront s'accorder sur leurs sources, et l'option d'introduire directement des données wikidata ne peut conduire qu'à une détérioration de la qualité de l'encyclopédie et au gaspillage de ressources humaines » (Wikipédia, 2017). Des propos similaires peuvent également être relevés dans le cadre des discussions menées par la communauté de Wikipédia en anglais, à l'instar de cette intervention laconique de l'utilisateur *Courcelles* : « Wikidata is useful for someone, somewhere, but not us. We should keep local control of all our information, rather than farm out the work to a project whose long-term viability I'm nowhere near convinced of. » (Wikipedia, 2019).

Rentrer dans une synthèse approfondie des divergences d'opinion vis-à-vis des données issues de Wikidata dépasserait cependant le périmètre de recherche de cette thèse, nous nous contenterons donc ici de renvoyer vers la synthèse résumant les différents arguments avancés par les membres de Wikipédia en anglais : Wikipedia (2019).

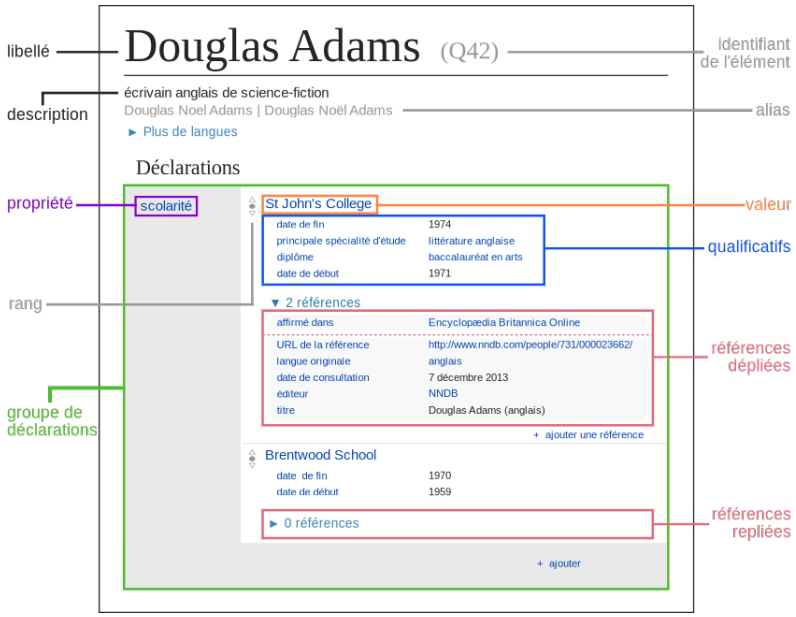


FIGURE 2.2 – Schéma détaillant les composants du modèle de données Wikidata, ici pour l'élément Q42 correspondant à l'écrivain britannique Douglas Adams. Source : Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Datamodel_in_Wikidata_fr.svg).

sance commence à prendre son essor [18]. Alors qu'au départ les éditeurs avaient seulement le loisir de créer des éléments et de les connecter à des articles Wikipédia, le modèle de données se développe afin de pouvoir stocker des données structurées au-delà de simples libellés et liens *interlangues* (Vrandečić et Krötzsch, 2014). La figure 2.2 est utilisée par Wikidata [19] pour expliciter ce modèle de données, en prenant l'exemple d'une déclaration associée à l'élément Q42|Douglas Adams [20]. Nous détaillons ces différents composants, étant donné qu'ils sont également utilisés dans le cadre de notre étude de cas présentée en seconde partie.

La figure 2.2 nous montre un extrait de la page web associée à une entité Wikidata. Cette entité fait partie des plus de 85 millions d'entités [21] constituant la base de connaissance Wikidata (MediaWiki, 2020). Cette entité – dans le cas présent un élément (*item* en anglais), mais cela pourrait également être une propriété, comme nous le verrons –, possède un iden-

18. Essor que favorise notamment l'import de données Freebase que Google décide d'offrir à Wikidata en 2014 (Pellissier Tanon *et al.*, 2016).
19. Sur ces pages d'aide et de glossaire, voir par exemple : https://www.wikidata.org/wiki/Wikidata:Glossary/fr.
20. https://www.wikidata.org/wiki/Q42.
21. 87 090 313 éléments et 7 580 propriétés début juin 2020 (Wikidata, 2020b), voir également : https://rawgit.com/johnsamuelwrites/wdprop/master/properties.html.

tifiant numérique unique, ici : Q42. Elle est décrite à l'aide de texte lisible par des humains : un libellé (*label* en anglais), une description, ainsi que d'éventuels alias. Ces éléments textuels, destinés à permettre son identification et sa désambiguïsation, peuvent être encodés dans n'importe quelle langue supportée par le logiciel. Ils composent ce que Wikidata nomme *fingerprint* (MediaWiki, 2020). Dans le cas de la figure 2.2, ces données sont en français :

— Libellé : Douglas Adams
— Description : écrivain anglais de science-fiction
— Alias : Douglas Noel Adams ; Douglas Noël Adams

Viennent ensuite les déclarations (*statements* en anglais) qui décrivent cette entité. Sur le schéma seule une déclaration concernant la scolarité de Douglas Adams est visible, mais il faut imaginer que la page [22] se poursuit avec des dizaines d'autres déclarations concernant son genre, son pays de nationalité, son lieu de naissance ou encore une liste de ses œuvres notables. Ces déclarations sont systématiquement composées d'une propriété [23] et d'une valeur [24]. Dans le cas de la figure 2.2, cette paire est composée de :

— Propriété : scolarité
— Valeur : St John's College

Cette combinaison [élément + propriété + valeur], forme ce qui correspond au triplet [sujet + prédicat + objet] dans le contexte du Web sémantique. Mais le modèle de données utilisé par Wikidata se singularise à ce niveau-là : à ce triplet peuvent venir s'ajouter d'autres données, destinées à préciser ou nuancer une déclaration. En effet, il faut se rappeler que :

> One of [our] requirements is that « Wikibase will not be about the truth, but about statements and their references. » This means that in Wikibase we do not actually model the items themselves, but statements about them. We do not say that Berlin has a population of 3,5 M, we say that there is this statement about Berlin's population being 3,5 M as of 2011 according to the German statistical office. (MediaWiki, 2020)

Pour ce faire, la personne éditant l'élément dispose de deux types d'agencement : d'une part, des références visant à relier une affirmation à sa source :

22. https://www.wikidata.org/wiki/Q42.
23. À laquelle est associée un identifiant numérique précédé d'un *P*, ainsi qu'une page web décrivant l'entité, voir par exemple : https://www.wikidata.org/wiki/Property:P69.
24. Il faut noter qu'un certain type de valeur est attribué à chaque propriété lors de sa création : la valeur sera alternativement un élément de la base de connaissance, une date, des coordonnées géographiques ou encore une quantité ; pour une liste exhaustive, voir : https://www.wikidata.org/wiki/Help:Data_type/fr.

d'autre part, des qualificatifs permettant de nuancer ou de préciser l'information, le cas échéant. Comme l'explique l'un des fondateurs de Wikidata, « we thus arrive at a model where the property-value pairs assigned to items can have additional subordinate property-value pairs we call *qualifiers* (Vrandečić et Krötzsch, 2014) ». La figure 2.2 illustre cela : elle indique que l'un des lieux de scolarité de Douglas Adams fut St John's College, *de* 1971 *à* 1974, *selon* l'Encyclopædie Britannica. Ce faisant, les qualificatifs permettent l'encodage de « ternary [25] relations that elude the property-value model » (Vrandečić et Krötzsch, 2014). Nous y reviendrons.

Enfin, si une déclaration possède plusieurs valeurs (par exemple, la ville de Bruxelles possède différentes valeurs pour la propriété P1082|Population : l'une date de 2017, l'autre de 2018), il est possible de privilégier une valeur en lui attribuant le rang *préféré*, tandis que les autres garderont le rang *normal* qui leur a été attribué par défaut. A contrario, il peut arriver que des déclarations soient considérées « comme étant erronnées ou représentant des connaissances dépassées » [26], dans de tels cas, il pourra être utile de leur attribuer le rang *obsolète*.

Alors que l'ensemble de ces informations, de même que l'intégralité du contenu de la base de connaissance, est stocké en interne au format JSON [27], Wikimedia a décidé de l'encoder également au format RDF [28] et de le stocker dans un *triple store* Blazegraph, « [to] address Wikidata's need to share, query, and analyse data in a uniform way » (Malyshev *et al.*, 2018).

Les figures 2.3 et 2.4 montrent cela : la première illustre comment, quelle que soit la façon d'ajouter des données – à travers une interface utilisateur ou à travers une API *(Application Programming Interface)* –, ces dernières sont stockées sous forme d'objets JSON par le biais d'une base de données SQL Shorland (2019c). La seconde figure montre comment les données JSON contenues dans la base de données SQL servent de base au fonctionnement du Wikidata Query Service. Le fonctionnement courant [29] de ce dernier consiste en une mise à jour en temps réel des triplets – encodés selon la syntaxe Turtle – dans le *triple store* Blazegraph, à chaque modification d'une entité Wikidata (Shorland, 2019c).

25. Notons toutefois que Erxleben *et al.* est plus réservé quant à l'utilisation de cet épithète : « Arguably this is a ternary relationship, but the boundary between context annotation and n-ary relation is fuzzy. For example, Star Trek : The Next Generation (Q16290) has cas member (P161) Brent Spiner (Q311453) with two values for qualifier character role (P453) : Data (Q22983) and Lore (Q2609295). » (Erxleben *et al.*, 2014).

26. https://www.wikidata.org/wiki/Help:Ranking/fr.

27. Le fichier est construit autour des entités Wikidata (éléments ou propriétés), qui constituent des objets JSON individuels. Le détail sur la structuration de ces données est repris sur cette page : https://www.mediawiki.org/wiki/Wikibase/DataModel/JSON.

28. Pour obtenir des détails sur ce processus et les questions que cela a soulevé, voir (Erxleben *et al.*, 2014).

29. Il existe également une option d'importation de *dumps RDF*.

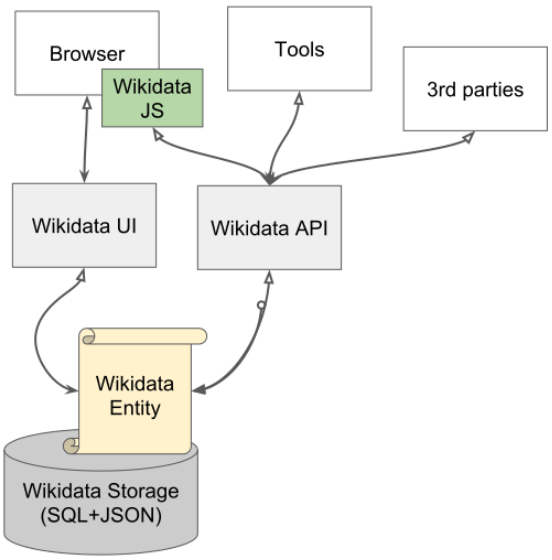


FIGURE 2.3 – Wikidata Architecture Overview - Data getting into Wikidata. *Source* : Addshore (CC BY-SA) (https://commons.wikimedia.org/wiki/File:Wikidata_Architecture_Overview_-_High_Level.svg).

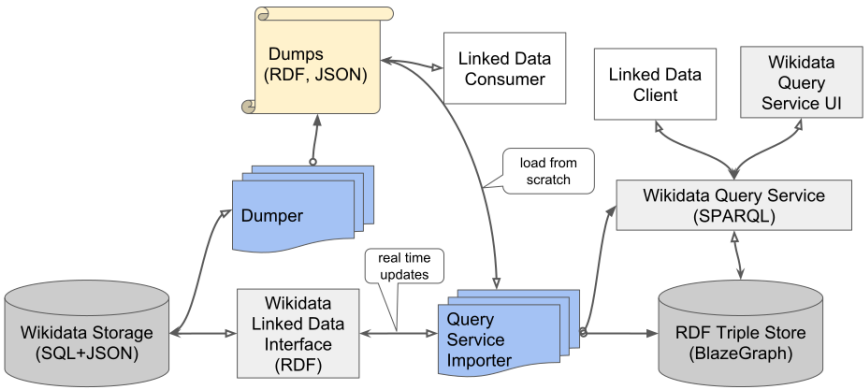


FIGURE 2.4 – Wikidata Architecture Overview - Query Service. *Source* : Addshore (CC BY-SA) (https://commons.wikimedia.org/wiki/File:Wikidata_Architecture_Overview_-_Query_Service.svg).

S'ils affirment que « conceptually, the graph-like structure of Wikidata is already rather close to RDF », Malyshev *et al.* relèvent toutefois que les composants de la base de connaissance Wikidata contiennent plus d'informations que ce que le RDF ordinaire ne peut prendre en charge [30]. L'encodage nécessite dès lors des ajustements :

> Complex values and annotated statements both are represented by auxiliary nodes, which are the subject of further RDF triples to characterise the value or statement annotations, respectively. (Malyshev *et al.*, 2018, p. 3)

Ce processus, connu sous le terme de réification [31], conduit à une multiplication des données :

> Overall this encoding of statements leads to graphs with many, sometimes redundant triples. This design is meant to simplify query answering, since users can easily ignore unwanted parts of this encoding, allowing queries to be as simple as possible and as complicated as needed. (Malyshev *et al.*, 2018, p. 4)

Bien qu'ils reconnaissent qu'une vue simplifiée puisse être souhaitée dans certains cas et qu'ils aient dès lors prévu à cet effet des exports RDF simplifiés (« not faithful but still meaningful »), Erxleben *et al.* arguent que cela ne suffit pas : « we cannot avoid to use relatively complex RDF graphs if we want to capture the rich structure of Wikidata statements » (Erxleben *et al.*, 2014, p. 12).

Cette complexité, couplée à l'usage d'une ontologie opaque et sur mesure [32], soulève cependant des réserves parmi la communauté du Web sémantique. Sans s'appesantir sur cette question qui dépasse le cadre de cette thèse, notons toutefois que certains s'interrogent sur l'impact de cette singularité en matière d'interopérabilité (Ismayilov *et al.*, 2018; Abián *et al.*, 2017;

30. En revanche, RDF* – une évolution de RDF, actuellement discutée par le W3C – vise également à permettre d'annoter des triplets RDF en proposant *an alternative approach that is based on nesting of RDF triples and of query patterns* (Hartig, 2019).

31. « The process of encoding complex structures in RDF by introducing new individuals to represent them », selon (Erxleben *et al.*, 2014, p. 10).

32. Paul Wilton, auteur de l'ouvrage « Beginning SQL » en dresse un portrait sans concession : « Instead of using the RDF schema to define the properties and classes, [W]ikidata have defined their own schema that sort of mirrors the RDF schema. There is a [W]ikidata property P31 "instance of" that is semantically equivalent to *rdf :type*. Property P279 "subclass of" is semantically equivalent to *rdfs :subClassOf*. Classes themselves are declared as items, and items are then described using these properties (not using the RDF schema). This abstraction layer, means [W]ikidata loses all the benefits of RDF schema, and gains a frankly painful amount of confusion. Any consumer of [W]ikidata first needs to understand the [W]ikidata schema. [...] The entire set of [W]ikidata properties including those that are the schema itself (the one that is not the RDF schema) are opaque and not human understandable. This makes building queries and applications on [W]ikidata really painful and time consuming. » (Wilton, 2018).

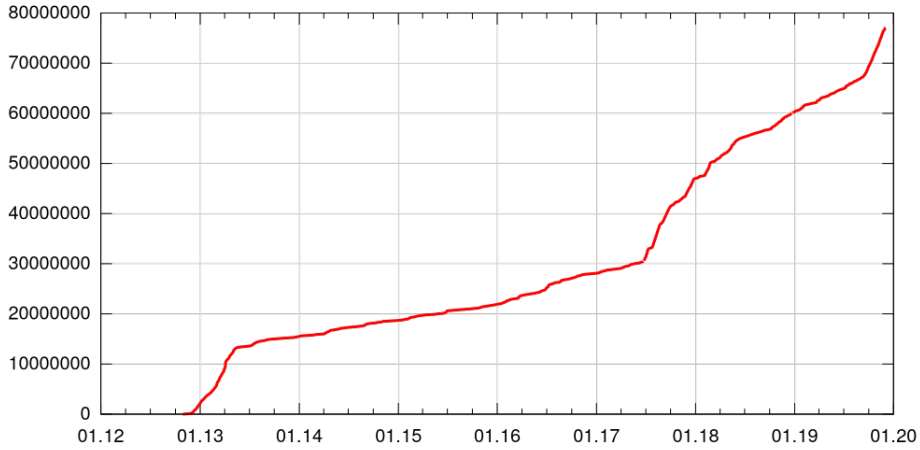


FIGURE 2.5 – *Plot of item creation by date for Wikidata* (décembre 2019). Source : Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Wikidata_item_creation_progress_no_text.svg).

Hitzler, 2020; Freire et Isaac, 2019; Färber *et al.*, 2018), d'interconnexion à des vocabulaires externes(Färber *et al.*, 2018; Erxleben *et al.*, 2014), mais également de formulation de requêtes SPARQL (Hernández *et al.*, 2015; Färber *et al.*, 2018) ou de facilité d'utilisation (Godby *et al.*, 2019; Färber *et al.*, 2018; Spitz *et al.*, 2016). Enfin, au-delà des limites propres au format des triplets qu'elle génère, Wikidata reste soumise aux mêmes challenges que les autres acteurs du Web de données lorsqu'il s'agit d'exprimer des réalités complexes et nuancées dans un format lisible par des machines (Brown et Simpson, 2013; Rizza *et al.*, 2019).

Pour revenir aux origines de Wikibase, il faut relever que bien que les efforts de Wikidata aient d'abord été mis au service de Wikipédia, la base de connaissance fut en réalité envisagée de façon plus large dès le départ :

> In addition to the Wikimedia projects, the data is expected to be beneficial for numerous external applications, especially for annotating and connecting data in the sciences, in government, and for applications using data in very different ways. The data will be published under a free Creative Commons license. (Matthew, 2012)

Cette dimension d'usage externe s'est développée au fur et à mesure de la croissance de la base de connaissance, Wikidata ayant conquis son indépendance vis-à-vis des autres projets Wikimedia et ayant désormais la possibilité d'accueillir de nouvelles pages sans que celles-ci soient forcément liées à des pages Wikipédia. Comme le montrent les figures 2.5 et 2.6, tant le nombre d'éléments créés que le nombre d'utilisateurs actifs par jour n'ont cessé de

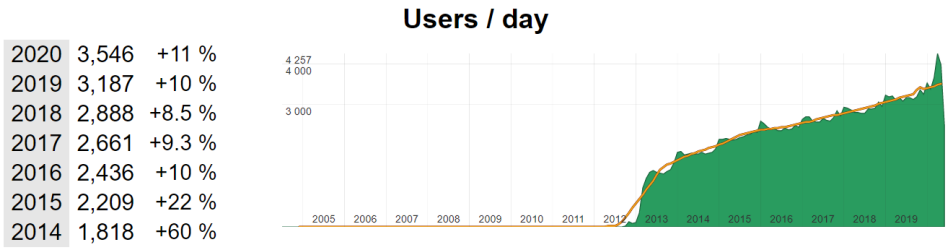


FIGURE 2.6 – Utilisateurs actifs sur Wikidata par jour, de 2005 à 2020. Source : Wikiscan (CC BY-SA) (https://wikidata.wikiscan.org/).

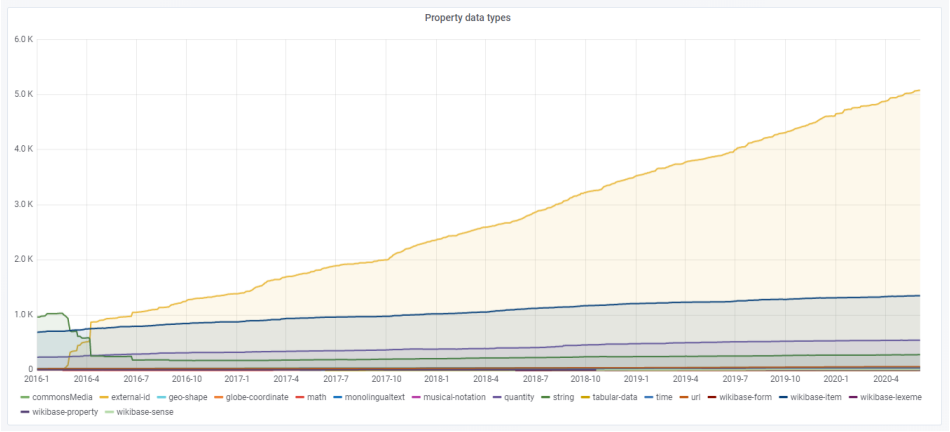


FIGURE 2.7 – Évolution des propriétés Wikidata par type de données, de janvier 2016 à juin 2020. Source : Grafana (https://grafana.wikimedia.org).

croître depuis le lancement de Wikidata. Au moment de la rédaction de cette thèse, Wikidata contient plus de 87 millions d'éléments – dont près de 10% se rapportent à des êtres humains –, elle a fait l'objet de près de deux milliards de modifications depuis le lancement du projet et elle compte plus de 25 000 utilisateurs actifs (Wikidata, 2020b).

Cette expansion a été accompagnée par une progression des propriétés Wikidata utilisant comme valeurs des identifiants externes *(external identifiers)* ; un *datatype* apparu sur Wikidata en 2016 (Wikidata, 2020d) afin de pouvoir traiter de façon adéquate ces identifiants issus de systèmes externes, qui étaient traités jusque-là comme des chaînes de caractères. Comme le montre la figure 2.7, ils ont connu une croissance continue, la barre des 5 000 ayant été franchie en mai 2020. Ces propriétés visant à stocker des identifiants externes composent ainsi à l'heure actuelle plus de 65 % de toutes les propriétés Wikidata[33].

33. Soit 5 078 sur un total de 7 580 propriétés.

L'outil de visualisation The Wikidata Identifiers Landscape [34], permet de voyager à travers cette *galaxie* [35] d'identifiants externes. Ainsi, il apparaît qu'en avril 2019, c'est l'identifiant P214|VIAF [36] qui possède le plus de chevauchements [37] avec d'autres identifiants externes. La figure 2.8, qui représente un réseau d'identifiants externes partageant la caractéristique d'être une propriété Wikidata pour notice d'autorité de personnalités [38], confirme le rôle central qu'y joue l'identifiant VIAF – en bas à gauche –, aux côtés d'autres *clusters* constitués autour de l'identifiant Freebase [39], de l'identifiant de la Bibliothèque du Congrès [40], de l'identifiant de l'Internet Movie Database [41] ou encore de l'identifiant Sports Reference [42] ; les identifiants périphériques étant principalement constitués d'identifiants associés à des sportifs, comme par exemple l'identifiant Tennis Archives [43].

Cette augmentation continue d'identifiants externes, encouragée notamment dans le secteur culturel (Europeana, 2017), s'est accompagnée d'une évolution dans la perception de Wikidata. Ainsi, Smith-Yoshimura relève au cours de l'été 2018 la place prise par Wikidata dans les réponses à l'étude internationale de l'Online Computer Library Center (OCLC) sur l'implémentation des données liées dans les bibliothèques : « Wikidata became the #5 ranked data source consumed by linked data projects/services described in the 2018 survey, compared to a #15 ranking in the 2015 survey » (Smith-Yoshimura, 2018b), tandis que Heberlein constate que « the GLAM [Galleries, Libraries, Archives, and Museums] community has rallied around Wikidata's cross-disciplinary potential » (Heberlein, 2019, p. 2). Le rapport du projet Passage, mené par l'OCLC, relève également qu'à mesure que le projet progressait « Wikidata was growing in importance to the library community as measured by the number of relevant presentations at professional conferences in 2017 and 2018 » (Godby *et al.*, 2019, p. 14). Cette reconnaissance du potentiel de Wikidata pour la gestion d'identités multilingues et de vocabulaires contrôlés (Heberlein, 2019; Bartholmei *et al.*, 2016) a en effet conduit à l'émergence [44] de la conception de Wikidata comme un

34. https://wmdeanalytics.wmflabs.org/WD_ExternalIdentifiersDashboard/.
35. Nous reprenons la métaphore utilisée par Wikidata : https://twitter.com/wikidata/status/1117689588485636096.
36. https://www.wikidata.org/wiki/Property:P214.
37. Ce chevauchement est basé sur le nombre d'identifiants *voisins* que possède l'identifiant VIAF sur chaque élément Wikidata qu'il identifie.
38. *Wikidata property for authority control for people* : https://ww.w.wikidata.org/wiki/Q19595382.
39. https://www.wikidata.org/wiki/Property:P646.
40. https://www.wikidata.org/wiki/Property:P244.
41. https://www.wikidata.org/wiki/Property:P345.
42. https://www.wikidata.org/wiki/Property:P1447.
43. https://www.wikidata.org/wiki/Property:P3670.
44. Comme le relèvent Allison-Cassin et Scott, cette idée est déjà présente en filigrane dans cet article publié en 2013 : *VIAFbot and the Integration of Library Data on Wikipedia* (Klein et

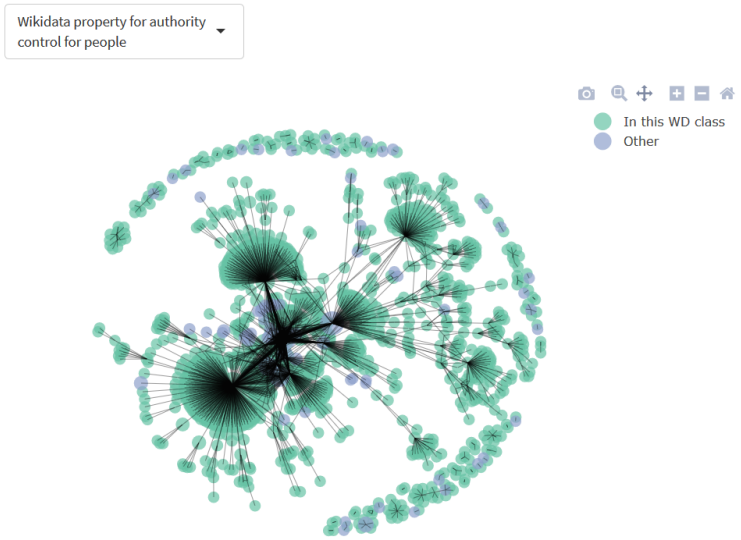


FIGURE 2.8 – Visualisation – basée sur l'algorithme Fruchterman-Reingold – du réseau d'identifiants externes Wikidata faisant partie des *propriétés pour notice d'autorité de personnalités* et de leurs voisins les plus proches (au niveau du nombre d'éléments en commun), en octobre 2019. Source : The Wikidata Identifiers Landscape, Goran S. Milovanovic - WMDE (https://wmdeanalytics.wmflabs.org/WD_ExternalIdentifiersDashboard).

hub central (Neubert, 2017; van Veen, 2017) ; une perception qui participe au renforcement du rôle joué par Wikidata dans le secteur du patrimoine culturel [45].

C'est dans ce contexte que Wikimedia Deutschland va saisir l'opportunité de donner forme à son rêve d'écosystème Wikibase, dans l'optique de délester Wikidata des données spécialisées qui n'y auraient pas leur place [46], tout en répondant à une demande croissante d'institutions souhaitant disposer de leur propre instance Wikibase (Pintscher *et al.*, 2019b).

Kyrios, 2013), bien qu'à l'époque Wikidata ne disposait pas encore de propriétés dédiées aux identifiants externes en tant que tels.

45. Elle est désormais identifiée « as a strategic direction by institutions including the International Federation of Library Associations and Institutions (IFLA), the Program for Cooperative Cataloging (PCC), the Association of Research Libraries (ARL), OCLC, and Linked Data for Production (LD4P) » (Heberlein, 2019, p. 2).

46. Waagmeester *et al.* suggère ainsi d'utiliser un critère voulant que le nombre total de concepts propres à un certain domaine ne devrait pas excéder 20% de Wikidata, il donne ainsi l'exemple des données relatives aux galaxies : « [Wikidata] is a general knowledge base and populating this with, for example, all known galaxies would instantly make it more a knowledge base on our universe, than a knowledge base that fits, which it currently is. The most conservative estimate of existing galaxies is about 200 million, while currently, the Wikidata contains 50 million items. Yet, there are compelling reasons to add knowledge about galaxies or similar size domains to a linked database » (Waagmeester *et al.*, 2018a).

2.1.2 Évolution

Cette sous-section s'intéresse à l'histoire de Wikibase et de sa communauté de développeurs et d'utilisateurs. Cette évolution est abordée à travers deux dimensions distinctes : le développement du logiciel d'une part ; l'émergence d'une communauté de l'autre. Avant d'explorer les possibilités et les limites de Wikibase en détail, le but ici est d'ébaucher l'évolution de cette suite logicielle en survolant les avancées de son développement, mais également en considérant comment l'importance accordée à Wikibase par les personnes qui développent Wikidata a évolué.

Il faut tout d'abord souligner qu'il n'est pas forcément aisé de dresser un historique, s'agissant d'un outil se développant en temps réel, en parallèle de cette thèse, et sur lequel nous ne disposons que de peu de recul. Il est néanmoins possible, en utilisant un ensemble de sources allant des rapports de l'équipe de développeurs Wikibase à la publication de nouvelles offres d'emploi, d'esquisser à grands traits les jalons ayant marqué cette évolution. Nous distinguons ainsi trois phases : les prémices, de la première utilisation du logiciel Wikibase hors Wikidata jusqu'au lancement de la première version *dockerisée*[47] du logiciel ; la consolidation, avec la résolution de bugs et l'adaptation du logiciel à un cercle plus étendu ; l'expansion, avec le développement des ressources allouées à Wikibase pour mieux servir l'objectif d'un écosystème Wikibase.

Pour commencer, il y a la phase de prémices. Cette dernière laisse déjà envisager la diffusion à venir de Wikibase, mais sans que cette dernière ne soit déjà une priorité. Au cours des premières années de développement de la base de connaissance Wikidata, le logiciel Wikibase n'est donc pas encore explicitement conçu pour un usage hors de Wikidata[48]. Il y a bien quelques pionniers qui s'aventurent déjà à réutiliser le logiciel à leurs propres fins, à l'instar de l'EAGLE project[49], un projet mené par Europeana, « the first production user of Wikibase outside Wikimedia, years before anyone else »[50], ou de Rhizome[51], qui utilise l'outil depuis 2015 (Pintscher *et al.*, 2019b). Il faut toutefois attendre 2016-2017 pour que des efforts soient entrepris afin de faciliter la réutilisation de Wikibase. Si ces quelques mots postés en janvier 2016 laissent présager de la suite : « to increase the ease of installation,

47. Cet adjectif, dérivé du nom du logiciel Docker (un logiciel libre permettant de lancer des applications dans des conteneurs logiciels), renvoie au fait qu'il s'agit d'une version Wikibase qui est mise à disposition dans un conteneur (Docker) contenant à la fois l'application et ses dépendances ; nous y reviendrons.

48. « Its initial purpose was powering Wikidata » (Pintscher *et al.*, 2019b, p. 3).

49. https://wiki.eagle-network.eu/wiki/Main_Page.

50. Installé en mai 2013 déjà, à en croire l'historique de l'installation : https://wiki.eagle-network.eu/wiki/Main_Page.

51. https://artbase.rhizome.org/wiki/Main_Page.

and hopefully increase number of installations/users/developers, it would be useful to include a Wikibase repository preconfigured »[52], c'est en 2017 que cette ambition prend véritablement forme.

Lors du cinquième anniversaire de Wikidata, en octobre 2017, une annonce discrète figure au bas de la liste des *présents* offerts à la communauté d'utilisateurs[53] : « From labs VM [Virtual Machine] to Wikibase Query Service in 2 minutes (using Docker images and docker-compose) »[54]. Ce sont les mots de l'un des développeurs Wikidata, Adam Shorland, plus connu sous son nom d'utilisateur, *Addshore*. Il explique brièvement sur son blog que ces images Docker devraient permettre « to quickly create docker containers for Wikibase backed by MySQL and with a SPARQL query service running alongside updating live from the Wikibase install » (Shorland, 2017). Si la consultation du dépôt GitHub[55] montre que les premiers efforts pour proposer cette version Docker remontent au printemps 2017 déjà – à l'occasion de la conférence WikiCite 2017 –, il prévient toutefois les potentiels intéressés :

> I wouldn't really recommend running any of these in *production* yet as they are new and not well tested. Various things such as upgrade for the query service and upgrades for mediawiki / wikibase are also not yet documented very well. (Shorland, 2017)

Cela n'empêche pas quelques esprits curieux de tester avec enthousiasme cette installation... Ainsi, quelques mois plus tard, Matt Miller, investi dans le Semantic Lab at Pratt, dédie un tutoriel à l'installation d'une instance Wikibase *for research infrastructures*, tutoriel qu'il illustre à l'aide de données issues du projet *Linked Jazz*, expliquant que le fichier d'instruction Docker *docker-compose.yml* se charge d'installer tout ce qui est nécessaire, de la base de donnée MySQL à la base de données Blazegraph (Miller, 2018). Son post est suivi par celui de Bob DuCharme, auteur de l'ouvrage *Learning SPARQL*[56], qui conclut ainsi :

> Many of us have waited years for an open-source framework that makes the development of web-based RDF applications as easy as Ruby on Rails does for web-based SQL applications. The dockerized version of Wikibase looks like a big step in this direction. (DuCharme, 2018)

Cette première version *dockerisée* représente ainsi un tournant dans le développement de Wikibase : il ne s'agit plus seulement d'utiliser ce logiciel

52. https://phabricator.wikimedia.org/T123019.
53. https://www.wikidata.org/wiki/Wikidata:Status_updates/2017_10_30.
54. L'annonce est accompagnée d'un lien renvoyant vers un dépôt GitHub dédié : https://github.com/wmde/wikibase-docker.
55. https://github.com/wmde/wikibase-docker/commit/deaaf50100f47df7a745f1cf2b7dad03b8d39dfb.
56. http://www.learningsparql.com/.

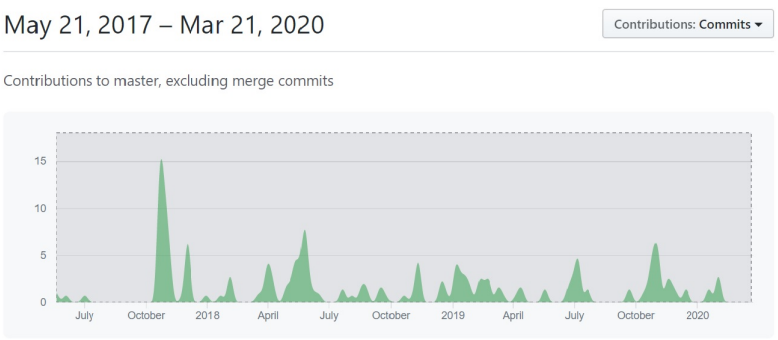


FIGURE 2.9 – Évolution dans le temps des contributions au dépôt GitHub *Wikibase-Docker*, du 21 mai 2017 au 21 mars 2020. Source : GitHub (https://github.com/wmde/wikibase-docker/graphs/contributors?from=2017-05-21&to=2020-03-21&type=c).

dans le cadre de Wikidata, mais de faciliter son appropriation par des tiers. Ce changement de positionnement stratégique[57] marque le passage de la phase de prémices à la phase de consolidation.

Comme le montre la figure 2.9, qui donne à voir l'évolution dans le temps des contributions *(commits)* au dépôt GitHub *Wikibase-Docker*, une fois passé le pic d'activité du mois d'octobre 2017 – coïncidant avec la démonstration proposée lors de la première conférence Wikidata[58] –, l'activité est continue, le logiciel faisant l'objet d'améliorations constantes. En témoigne également le nombre de *tâches* dédiées sur Phabricator[59] – la plateforme collaborative de Wikimedia utilisée entre autres pour le développement et le suivi logiciel.

Ainsi, bien que la plateforme n'offre pas de statistiques détaillées[60], elle permet toutefois de suivre les efforts en cours, qu'il s'agisse de la résolution de bugs ou du développement de nouvelles fonctionnalités. Par ailleurs, un export de toutes les *tâches* (ouvertes ou clôturées) associées au projet *Wikibase-Containers*[61], permet d'extraire leur date de création et ainsi de connaître la distribution dans le temps de ces 212 tâches, reprises dans le tableau 2.1. Le pic observé en 2018, avec l'ouverture de 114 tâches, illustre bien la dimension de consolidation de cette seconde phase : le nombre d'uti-

57. « In 2017/18 we added the idea of an ecosystem around Wikidata as an important next step because we believe that this is a key step to spreading more knowledge. » (Pintscher *et al.*, 2019a, p. 5).
58. https://www.wikidata.org/wiki/Wikidata:WikidataCon_2017.
59. https://phabricator.wikimedia.org/.
60. « Phabricator doesn't support further statistics, metrics, charts, reports (e.g. over time) or however you may call them, whether built-in or via an API » (MediaWiki, 2020b).
61. Il s'agit donc des *containers* Docker permettant d'installer plus facilement le logiciel Wikibase, voir : https://phabricator.wikimedia.org/project/view/3079/.

Année	**Tâches**
2016	1
2017	13
2018	114
2019	60
2020	24

TABLE 2.1 – Récapitulatif du nombre de tâches *Phabricator* associées au projet *Wikibase-Containers* par année, de début 2016 au 15 avril 2020.

lisateurs intéressés par Wikibase croît, entraînant une hausse du nombre de questions, de signalements de dysfonctionnements ou encore de demandes de nouvelles fonctionnalités. Comme le relève les développeurs Wikidata :

> When it became easier to set up a Wikibase instance, interest grew even more. [...] In 2018 alone, 11 new instances were set up. We believe this increase comes from having created a mature product as well as from investing in the Wikibase community and supporting potential users.(Pintscher *et al.*, 2019b, p. 8)

Outre ces données statistiques, l'importance croissante accordée à Wikibase transparaît également dans ce message de l'équipe de développement de Wikidata, posté en octobre 2018 à l'occasion du sixième anniversaire de la base de connaissance :

> Since we started the development of Wikidata we always had it in the back of our minds that Wikidata isn't the only place where Wikibase, the software running Wikidata, will be used. We always imagined other Wikibase instances out there next to Wikidata. However we didn't have the resources to focus on it and to be honest the rest of the world wasn't ready for it yet either. This has changed now. Wikidata has grown and more and more institutions, companies and projects are interested in not just contributing to Wikidata but also setting up their own knowledge base to open up their data. So over the last year we got the wheels in motion to create an ecosystem of Wikibase installations around Wikidata. (Pintscher, 2018)

Ces changements stratégiques [62] s'accompagnent de mesures très concrètes, telles qu'un partenariat mené par Wikimedia Deutschland avec la Biblio-

62. La volonté de créer un écosystème Wikibase est par exemple intégré aux axes stratégiques de la nouvelle stratégie pluriannuelle de Wikimedia Deutschland initiée en septembre 2018 (Wikimedia Deutschland, 2019a).

thèque nationale allemande (Ohlig, 2018), qui sera détaillé au cours des pages suivantes.

Le but ici n'est pas de pousser l'exercice de ce survol historique jusqu'à l'identification d'une borne temporelle précise qui marquerait le passage de la phase de consolidation à celle de l'expansion. L'année 2019 représente sans doute une zone grise à cheval entre les deux. D'une part, des efforts soutenus sont menés pour améliorer l'utilisabilité du logiciel : de nouvelles *images Docker* sont lancées, au gré du développement du logiciel MediaWiki [63] ; plusieurs améliorations du code [64] et de la documentation voient le jour, notamment au cours du Wikimedia Hackathon 2019 (Wikibase Community, 2019) ; en octobre WBStack [65] – un service permettant de mettre en place des instances Wikibase sans connaissances techniques – est lancé à l'occasion de la conférence WikidataCon 2019 (Shorland, 2019a), etc. D'autre part, des ambitions plus élevées se profilent, révélées notamment à travers la parution simultanée de quatre rapports stratégiques par la Wikimedia Foundation (WMF) et Wikimedia Deutschland (WMDE) au cours de l'été 2019.

Ces rapports [66], publiés sept ans après les premières étapes de développement de Wikidata, témoignent d'une volonté de passer à l'étape suivante, à savoir la phase de dissémination de Wikibase. Composés d'un *product vision paper* et de trois *product strategy papers*, ils visent à répondre à un besoin de clarté exprimé par les personnes intéressées par Wikidata et Wikibase. Le rapport intitulé *Strategy for the Wikibase Ecosystem* fournit d'utiles renseignements. Il met notamment en lumière la volonté d'assurer la création de nouveaux produits et services, en tirant parti de l'opportunité qui se présente :

> In the 2000s, the tech world saw an explosion in mapping APIs which brought together disparate but powerful data [...] We have the opportunity to be at the forefront of a new but similar phenomenon – linked data services that can power highly useful products we can't even imagine today. (Pintscher *et al.*, 2019b, p. 6)

Saisir cette opportunité requiert de travailler sur différents fronts : que ce soit en fournissant de meilleures possibilités techniques d'interconnexion de différentes bases de connaissance ou en nouant de nouveaux partenariats, mais également en se prémunissant des risques menaçant le succès d'une telle entreprise, à savoir le risque de manquer l'opportunité de croissance

63. Qui fait partie intégrante de la suite logicielle Wikibase.
64. Comme par exemple l'intégration de QuickStatements, qui était problématique jusque-là.
65. https://www.wbstack.com/.
66. Voir : https://meta.wikimedia.org/wiki/Wikidata/Strategy/2019.

qui se présente et celui de perdre l'occasion de créer un réseau de données liées ouvert [67]. Or, l'organisation fait face à différentes limites :

> At the same time we are reaching the limits of our technical, organisational and social infrastructure. [...] We are now at a strategic decision point and need to consciously decide where to take Wikidata and Wikibase in the future and what their roles are in relation to Wikipedia and the other Wikimedia projects. We have amazing opportunities in front of us but in order to fulfill them we need to scale our processes and organisational set-up. (Pintscher *et al.*, 2019a, p. 7)

Cette prise de conscience s'accompagne de mesures très concrètes, comme l'engagement de deux nouveaux collaborateurs par le département Software and Development de Wikimedia Deutschland [68], démontrant le fait que Wikimedia Deutschland considère désormais Wikibase comme un projet à part entière, distinct de Wikidata (Wikibase Community, 2019) ; ou encore la priorité accordée au développement de certaines fonctionnalités très attendues telles que la fédération de différentes instances Wikibase (Wikidata, 2020b).

S'il est encore trop tôt pour discerner si ces efforts suffiront à combler les limites auxquelles sont confrontés tant l'équipe de développement de Wikibase que ses premiers utilisateurs, il est en en revanche d'ores et déjà possible de dresser un tableau de ces limites, mais également des possibilités offertes par ce logiciel – ce que nous nous proposons de faire dans la section suivante, après un détour du côté de la communauté d'utilisateurs Wikibase.

Comme nous l'avons vu, la création d'un dépôt GitHub *Wikibase-Docker* – permettant une installation facilitée de Wikibase – remonte au mois de mai 2017 [69]. Il faut toutefois attendre février 2018 pour que soit officiellement créé un *Wikibase Community User Group* (Wikibase Community, 2018). Ce dernier est destiné à fournir aux utilisateurs un support technique plus accessible que ne l'était la plateforme Slack utilisée jusque-là. Bien que la communauté se dote de différents canaux de communication [70], les échanges restent globalement peu fréquents, en dehors du groupe de conversation Telegram qui connaît un succès croissant, comme nous le verrons. Ainsi, entre

67. « If we don't keep up and extend our game, this is likely to be taken over by other players and we will lose the ability to steer the development. [...] If we don't invest in Wikibase now, there is a chance commercial players will take over, and the data network will not be free and open. » (Pintscher *et al.*, 2019b, p. 8).

68. Un(e) *community communication manager*, ainsi qu'un(e) *partner relationship*, voir : https://lists.wikimedia.org/pipermail/wikibaseug/2020-January/000055.html. ; https://lists.wikimedia.org/pipermail/wikibaseug/2020-January/000056.html.

69. Le premier *commit* date du 25 mai 2017.

70. *Mailing list*, canal Slack, canaux IRC, groupe de conversation Telegram.

le moment de notre inscription à la *mailing list* – en août 2018 –, et le dernier envoi en date – au moment de rédiger ces lignes, c'est-à-dire en mars 2020 –, seuls 24 fils de conversation ont vu le jour [71], alors même que 137 personnes y sont inscrites [72].

En revanche, le groupe Telegram [73], qui se caractérise par des échanges plus informels mettant en contact direct utilisateurs et développeurs, connaît lui une plus grande activité. Comptant 12 membres 24 heures après sa création, en novembre 2018, 42 membres en mars 2019 et 143 membres en mars 2020, il est considéré aujourd'hui comme l'épine dorsale du groupe d'utilisateurs (Wikibase Community, 2020). Bien que ce groupe Telegram s'impose comme le canal de communication privilégié, le rapport annuel 2019 fait également mention des fréquentes interactions prenant place sur le réseau social Twitter, basées sur l'usage du hashtag #Wikibase [74]. Par ailleurs, il apparaît que certaines personnes se servent – par ignorance, ou peut-être sciemment, dans l'espoir d'obtenir davantage de réactions ? – de la *mailing list* Wikidata pour s'entretenir de sujets concernant Wikibase [75]. Enfin, une attention particulière est accordée à l'organisation de *community calls* depuis 2020 [76].

Parmi les jalons qui ont marqué le développement d'une communauté de pratique Wikibase, nous pourrions citer les trois workshops internationaux organisés au cours de l'année 2018 – à Anvers en avril 2018, à Berlin en juillet 2018 et à New-York en septembre 2018. La première rencontre, à Anvers, visait à explorer la possibilité de créer « a federated landscape of Wikibase insances federated with Wikidata » (Wikidata, 2018). Ce workshop a notamment abouti à la création d'un registre des instances Wikibase publiques [77]. Au mois d'août 2018, 18 instances étaient répertoriées ; en juin 2019, ce nombre s'élève à 26 ; en mars 2020 il en existe 30 [78]. Cependant, ces nombres sont à prendre avec de la distance : d'une part, parce que la liste

71. Les archives, contenant 17 *volumes* au mois de mars 2020, sont consultables en ligne : https://lists.wikimedia.org/pipermail/wikibaseug/.

72. En mars 2020.

73. Qui peut être rejoint via ce lien : https://t.me/joinchat/HGjGexZ9NE7BwpXzMsoDLA.

74. https://twitter.com/search?q="wikibase".

75. À l'instar de ces messages postés tant en 2018 : *Wikibase docker images and sitelinks*, qu'en 2020 : *Use case and thoughts on a local Wikibase with some simple* federation *with Wikidata*, voir : https://www.mail-archive.com/wikidata@lists.wikimedia.org/msg06215.html ; https://www.mail-archive.com/wikidata@lists.wikimedia.org/msg07540.html.

76. « The Wikibase community needs a venue for regular meeting and communication. We are coordinating calls on a regular basis to support this. » Wikidata (2020b) ; la première de ces réunions en ligne a eu lieu en février 2020 (Wikibase Community, 2020).

77. https://wikibase-registry.wmflabs.org/wiki/Main_Page.

78. Il est par ailleurs intéressant de mettre en balance cette progression avec celle de la période précédant l'arrivée d'une version Wikibase *dockerisée*, à savoir les années 2012 - 2017, au cours desquelles 12 instances seulement avaient été créées (Pintscher *et al.*, 2019b).

inclut des instances de test [79] et des instances tout à fait inactives [80] ; d'autre part, parce que toutes les instances Wikibase n'y sont pas répertoriées [81].

Ce type de rencontres internationales se poursuit en 2019, que ce soit sous la forme de workshops [82], dans le cadre de *meet ups* [83] ou de présentations dédiées à Wikibase dans le cadre de conférences internationales [84] ou encore lors de la journée d'étude *Linking the Past* organisée dans le cadre du projet Adochs [85].

Mais ce qui fait de 2019 une année-clé dans l'évolution du projet Wikibase [86], c'est surtout le fait que d'imposantes organisations, telles que la DNB (Deutsche Nationalbiblioteek), la BnF (Bibliothèque nationale de France), associée à l'Abes (Agence bibliographique de l'enseignement supérieur), ou encore l'OCLC (Online Computer Library Center) Research, commencent à partager leur expérience avec Wikibase pour la gestion de leurs données d'autorité et autres métadonnées bibliographiques.

En Allemagne, la DNB commence à s'intéresser à Wikibase en 2018 : l'outil est évoqué au mois de mai lors des Journées de l'Abes [87] par le responsable du GND (GND pour *Gemeinsame Normdatei*, c'est-à-dire *fichier d'autorité intégré*), Jürgen Kett [88]. En octobre, le projet *GND X Wikibase* est lancé (Ohlig, 2018), avant d'être évoqué plus publiquement lors de la GNDCon 2018, conférence placée sous le signe de l'ouverture des données [89] et du-

79. Comme par exemple *Test Wikimedia Commons*, qui annonce explicitement que l'instance sera supprimée à terme, après avoir servi à des tests dans le cadre du déploiement de *Structured Data on Commons* (voir https://test-commons.wikimedia.org/wiki/Main_Page).

80. Comme l'instance PlantData : http://wikibase-registry.wmflabs.org/wiki/Item:Q7.

81. C'est le cas par exemple de l'Eurhisfirm project https://wikibase.eurhisfirm.eu/wiki/Main_Page.

82. Par exemple à l'Université de Gand en juillet 2019https://www.wikidata.org/wiki/Wikidata:Events/UGent_Wikidata_and_Wikibase_Workshop_2019, lors de la conférence DCMI 2019 à Seoul : https://wikibase.peatix.com/ ou encore en lisière de la conférence Wikidata en octobre 2019 (https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Attend/Side_events/Wikibase_workshop_Berlin_2019).

83. Dans le cadre de Wikimania 2019 https://wikimania.wikimedia.org/wiki/2019:Meetups/Wikibase_meetup ou de la WikidataCon 2019 https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Program/Sessions/Wikibase_meetup.

84. Par exemple lors de la Knowledge Graph Conference 2019 (https://www.knowledgegraph.tech/speakers/ron-snyder/.), lors de la conférence LD4 2019 (https://wiki.lyrasis.org/display/LD4P2/2019+LD4+Conference+on+Linked+Data+in+Libraries), de la WikidataCon 2019 (https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Program/Sessions/Wikibase_inspiration_panel).

85. https://www.wikidata.org/wiki/Wikidata:Events/LkPast-WB.

86. https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Program/Sessions/Wikibase_inspiration_panel.

87. http://www.abes.fr/Publications-Evenements/Journees-Abes/Journees-ABES-23-24-mai-2018.

88. Qui précise quelques mois plus tard sur Twitter : « We evaluate wikibase as a possible *second home* for the GND. » (@JuergenKett, 2018).

89. Comme en témoigne le titre de l'édition 2018 : *Öffnung der GND* (ouverture du GND), voir : https://wiki.dnb.de/display/GNDCON2018/GNDCon+2018.

rant laquelle Wikibase fut apparemment *sur toutes les lèvres* et même évoqué comme *sauveur* (@scholl_i, 2018). Sur son Wiki, la DNB explique que la stratégie pour le GND au cours des années à venir est de favoriser la collaboration et le travail collaboratif :

> We look for solutions that will support data hosting and maintenance across domains. Collaboration will be the core idea. Therefore, Wikibase appears to be an ideal candidate to enlarge the GND environment. Wikibase could become a second home for GND in order to offer applications for access and contribution to new target groups, that cannot be served by the specific interfaces which are currently offered. (Fischer, 2018)

Le projet se précise en 2019 : outre son partenariat avec Wikimedia Deutschland, la Bibliothèque nationale débute une nouvelle collaboration avec l'Université d'Erfurt [90], qui utilise Wikibase depuis 2017 dans le cadre du projet FactGrid [91]. La bibliothèque nationale crée ainsi trois instances Wikibase, destinées à modéliser le GND selon des perspectives différentes [92], de manière à proposer un système plus attractif pour les futurs contributeurs, mais également pour tester la synchronisation entre ces instances, la gestion des droits d'administration des données ou encore des outils de visualisation de données (Fischer et Ohlig, 2019).

Après plusieurs mois de tests, Jens Ohlig (Wikimedia Deutschland) et Barbara Fischer (DNB) proposent en mars 2020 un bilan à deux voix, destiné à évaluer la pertinence de Wikibase pour le GND, et notamment sa capacité à simplifier la collaboration entre les bibliothèques et d'autres communautés. Leur preuve de concept leur a permis d'évaluer différentes dimensions du logiciel [93] et de parvenir à des résultats encourageants :

> Wikibase is a bridge to the world of open, cooperative and interdisciplinary authority control data. The proof of concept shows that this bridge is viable. (Fischer et Ohlig, 2020)

Si la DNB va continuer à utiliser son système actuel pour la maintenance de ces fichiers d'autorité (Fischer et Ohlig, 2020), les résultats concluants de

90. https://blog.factgrid.de/archives/1527.

91. *The Wikibase instance for historians :* https://database.factgrid.de/wiki/Main_Page.

92. « The first database represents the GND as used and processed by libraries today. A second database models the GND extended by the additional needs of cultural institutions such as museums and archives. And finally, the third database, Factgrid, creates a research database for historical persons and corporations on the basis of GND data records, which will no longer be an actual authority file. » (Fischer et Ohlig, 2019).

93. « What could a modular *GND 2.0* in Wikibase look like that meets the requirements of the different sectors ? How can the rules for the modelled properties of the entity types be mapped effectively and clearly ? How can a stable synchronization between a GND Wikibase instance and the CBS-based master instance be implemented ? » (Fischer et Ohlig, 2020)

cette phase d'évaluation lui permettent de confirmer qu'une instance Wikibase fera office de *domicile secondaire* pour le GND :

> With Wikibase we want to create an extended access to the GND for interest groups for whom the librarian editorial interfaces are not suitable. (Fischer et Ohlig, 2020)

Comme nous le verrons au cours des pages suivantes, ce type d'expérience permet également de mettre en lumière les limites de ce logiciel et les adaptations qui pourraient être envisagées à l'avenir.

En France, la BnF (Bibliothèque nationale de France) et l'ABES (Agence Bibliographique de l'Enseignement Supérieur) envisagent d'utiliser Wikibase pour leur Fichier National d'Entités (FNE). Ce dernier[94] commence à être évoqué publiquement au printemps 2017[95]. Il s'inscrit dans le cadre d'un programme national de transition bibliographique se traduisant par un renouvellement des outils de production des métadonnées et par l'utilisation de nouvelles formes de modélisation des données[96].

Après une étude de faisabilité pilotée par l'Abes et la BNF, les choses se précisent en 2019 avec la réalisation d'une preuve de concept « fondée sur l'infrastructure de Wikibase afin de bénéficier de l'infrastructure logicielle et des outils d'alignement et de curation des données existants en open source » (ABES et BNF, 2020). Elles collaborent avec deux développeurs qui disposent de six mois pour « vérifier si Wikibase – ainsi que tous les mécanismes qui lui sont associés [...] – répondent bien aux besoins du projet FNE » (ABES, 2019).

Lors de la WikidataCon 2019, les porteurs du projet présentent des premières conclusions nuancées. Bien qu'ils relèvent que le logiciel ne présente pas d'obstacle majeur pour importer leurs données – selon une ontologie spécifique – et qu'il présente des fonctionnalités qui pourraient constituer la base de l'infrastructure technique du futur FNE, ils constatent toutefois un écart entre les besoins spécifiques de l'institution et les possibilités offertes par Wikibase (Angjeli et Bober, 2019). Ils précisent par ailleurs que l'éventuelle adoption de l'infrastructure Wikibase n'a pas encore été tranchée. Au moment de la rédaction de ces lignes – février 2020 –, le dépôt GitHub associé au projet est toujours librement consultable[97], en revanche, aucune

94. À l'époque, c'est un Fichier national *d'autorité* qui est évoqué.

95. Cependant, dans un dossier de la revue Ar(abes)ques consacré aux *Autorités, référentiels, entités*, on apprend que ce projet avait été jugé prioritaire par le Comité stratégique bibliographique en 2015 déjà (Johannic-Seta, 2017)

96. Le FNE constitue ainsi « une des concrétisations du programme national Transition bibliographique par la mise en oeuvre du modèle IFLA-LRM. À ce titre, son périmètre cible englobe bien plus que les données d'autorités *traditionnelles* en incorporant la plupart des entités définies par le modèle IFLA-LRM (Agents, Oeuvres, Concepts, Laps de temps, Lieux). » (ABES et BNF, 2020).

97. https://github.com/abes-esr/poc-fne/.

décision n'a été exprimée publiquement quant à l'éventuelle adoption Wikibase pour la réalisation effective du FNE [98].

Par ailleurs, il s'avère que la BnF a jeté une seconde fois son dévolu sur Wikibase pour réaliser une autre preuve de concept – indépendante de la première citée ci-dessus –, dans le cadre du projet NOEMI (pour *Nouer les Œuvres, les Expressions, les Manifestations et les Items*) cette fois-ci. Ce projet lancé en 2017 vise à doter la BnF d'un nouvel outil de production des métadonnées, lui permettant de s'adapter à « un contexte d'évolution des tâches de catalogage et des tâches associées » (BNF, 2018). Cette fois encore un marché public a permis à la Bibliothèque nationale de s'associer à deux informaticiens, chargés de tester au cours de l'été 2019 la capacité du logiciel Wikibase à satisfaire les besoins formulés dans le cadre du projet NOEMI. À nouveau, aucune annonce publique ne permet à l'heure actuelle de savoir si l'outil a été définitivement adopté, en dépit de certaines réserves [99] soulignées lors de la WikidataCon 2019 (Senalada, 2019).

Enfin, l'OCLC Research publie en août 2019 un rapport détaillé sur le projet Passage, lancé dans le cadre de son programme de recherche sur les *Linked Data* [100]. Ce projet de 10 mois avait pour objectif de fournir un prototype aux bibliothécaires de 16 institutions américaines afin qu'ils puissent s'essayer à la description de ressources à l'aide de *Linked Data* sans devoir bénéficier d'expertise technique (Godby *et al.*, 2019). Si ce projet est aujourd'hui clôturé, l'OCLC a annoncé qu'un autre prototype Wikibase allait être développé dans le cadre de la gestion des *Linked Data* associées à des collections numérisées (Proffitt, 2019, p. 10). Par ailleurs, il faut garder à l'esprit que plutôt qu'un produit fini, le projet Passage avait surtout pour ambition de fournir un *bac à sable* [101] donnant l'opportunité à une communauté de « [to] share work in progress, have philosophical and practical conversations, share inspirations and concerns » (Proffitt, 2019, p. 7). Ce qui semble avoir été atteint, à en croire la conclusion du rapport Passage :

> The Passage pilot represented an opportunity for all participants to gain hands-on experience creating structured data that could be exported as linked data. [...] The results of this effort will

98. Mise à jour, septembre 2020 : au cours du mois de juillet 2020, une synthèse sur cette phase d'expérimentation a été publiée, accompagnée de la confirmation que la preuve de concept « a permis de conforter l'analyse préalable qui avait abouti à pressentir que Wikibase était une option technique viable et intéressante », voir : https://www.transition-bibliographique.fr/2020-07-13-preuve-concept-fne-synthese-decisions/.

99. Ces dernières concernaient des questions de performance, l'absence de facettes et de vues filtrées dans le moteur de recherche, ainsi que des difficultés avec la base de données postgreSQL associée à la Wikibase (Senalada, 2019).

100. https://www.oclc.org/research/themes/data-science/linkeddata/linked-data-outputs.html

101. Bac à sable – de l'anglais *sandbox* – fait référence dans ce contexte à un environnement de test de logiciel.

> help materialize the paradigm shift that is evoked by the name of the pilot. The shared goal is a *passage* from standards introduced in the 1960s to a 21st century solution featuring structured semantic data that promises better connections between libraries and the world beyond. (Godby *et al.*, 2019, p. 74)

Mais l'engouement des bibliothèques ne se limite pas à ces quelques exemples. Ainsi, Europeana a profité de l'édition 2019 de la conférence anuelle Wikimania pour organiser une session à destination des bibliothèques nationales intéressées par Wikidata ou Wikibase. Cette *rencontre inaugurale* a rassemblé une trentaine de représentants institutionnels venus de trois continents, dont 60% ont affirmé qu'ils envisageaient l'utilisation de Wikibase au sein de leur institution (Byrne et Wyatt, 2019). Un intérêt souligné également dans le rapport *Strategy for the Wikibase Ecosystem* publié à la même période [102], dans le livre blanc de l'ARL (Association of Research Librairies) publié quelques mois plus tôt [103], ou encore dans les résultats préliminaires diffusés par la BnF [104].

Cet intérêt croissant pour Wikibase s'accompagne également du constat que davantage de collaboration semble nécessaire. Ainsi, Liam Wyatt, organisateur de cette rencontre inaugurale, remarque que les projets déjà menés autour de Wikidata et de Wikibase se déroulent généralement de façon isolée, limitant les possibilités de partage de connaissances (Byrne et Wyatt, 2019). Même son de cloche du côté de la BnF, qui constate l'émergence d'une volonté collective de « créer une Fédération qui respecte les us et coutumes de chaque pays » (Johannic-Seta et Aymonin, 2019, p. 8).

Enfin, il est clair qu'au-delà du domaine des bibliothèques, d'autres Wikibases ont essaimé en 2019, à l'image de PersonalData.io [105] – *the integrative toolbox addressing surveillance capitalism* –, ou de l'EU Knowledge Graph [106] – *[a] graph [which] contains structured information about the European Union*.

102. « Recently we have seen a lot of activity from the GLAM sector. [...] Seven national libraries, among them Germany and France, have run substantial pilots, and libraries from seven more countries have communicated interest in evaluating or using Wikibase or Wikidata as a platform for creating or participating in linked data work for the sector. » (Pintscher *et al.*, 2019b, pp. 3-4)

103. « Wikibase has a growing community of users in the GLAM and research sectors. [...] The growing number of Wikibase implementations suggests opportunities for scholarly and GLAM reuse of the software as a generic data store » (Allison-Cassin *et al.*, 2019, pp. 38-39)

104. « La veille internationale confirme : qu'il y a autour de Wikidata/Wikibase une communauté solide dans laquelle les bibliothèques s'impliquent de plus en plus ; qu'Israéliens, Luxembourgeois, Allemands, EURIG, laboratoires, bibliothèques s'intéressent à Wikibase » (Johannic-Seta et Aymonin, 2019, p. 8).

105. https://wiki.personaldata.io/

106. https://linkedopendata.eu/wiki/The_EU_Knowledge_Graph

Quant à l'année 2020, qui continue à voir de nouvelles bibliothèques se lancer dans des expérimentations faisant appel à Wikibase, à l'instar de la Bibliothèque nationale de Finlande [107], elle verra peut-être également éclore des projets du côté des archives. En effet, outre notre projet de Wikibase réalisé dans le cadre de cette thèse, nous n'avons pour l'instant croisé qu'une seule initiative issue du milieu archivistique, à savoir un prototype éphémère développé dans le cadre du premier hackathon des Archives nationales de France, qui s'est déroulé en décembre 2018. À cette occasion, les six membres de l'équipe *Wikibase Archives* avaient choisi de travailler sur le sixième défi proposé au cours du hackathon, à savoir « les clés d'accès à quatorze siècles d'histoire », et d'y répondre en créant une instance Wikibase destinée à accueillir tant les référentiels que les instruments de recherche de l'institution (Archives nationales de France, 2018).

Cependant, il ne faudrait pas oublier Rhizome, qui fait figure de précurseuse en étant l'une des premières organisations à avoir utiliser le logiciel Wikibase à ses propres fins (Fauconnier, 2018). En effet, bien qu'il ne s'agisse pas d'un centre d'archives au sens strict [108], il s'avère que l'organisation utilise Wikibase depuis 2015 déjà, dans le cadre de la mise en ligne de ses archives : « we use Wikibase in an experimental way, but we are committed to adopting it as a long-term and sustainable solution for the ArtBase » (Fauconnier, 2018). En parallèle des efforts effectués pour migrer le contenu de cette base de données décrivant plus de 2 000 œuvres d'art vers une instance Wikibase [109], l'équipe de Rhizome se concentre également sur l'expérience utilisateur, par exemple en œuvrant à la refonte de son interface de recherche afin de tirer parti de la puissance des requêtes SPARQL permises par Wikibase (Fauconnier, 2018). Cette dernière n'est pas encore en ligne actuellement : l'organisation a affirmé vouloir encore développer certaines fonctionnalités avant de passer à la mise en production complète de sa Wikibase [110].

Si les initiatives sont encore rares dans le secteur des archives, le sujet semble toutefois susciter de l'intérêt au sein de la communauté, à en croire certains échanges de tweets. Ainsi, au cours d'une discussion portant sur les

107. L'utilisateur *JarmoS (sarrit)* écrit le 20 février sur le groupe Telegram *Wikibase community* : « Someone asked about GLAMs. Just to let you know, at the Finnish National Library, we will be developing / testing an agent authority database for Finnish Glam sector on a WB. Probably close match with the German and French national libraries' efforts. » [Message Telegram] JarmoS (sarrit) 20.02.2020.

108. L'organisation est décrite comme « a not-for-profit arts organization that supports and provides a platform for new media art »(Wikipedia, 2020).

109. https://artbase.rhizome.org/wiki/Main_Page.

110. Message envoyé en octobre 2019 à la *mailing list* Wikibase Community User Group par Dragan Espenschied, *Rhizome's preservation director*, voir : https://lists.wikimedia.org/pipermail/wikibaseug/2019-October/000046.html.

technologies sémantiques et les éditeurs de systèmes de gestion d'archives, un archiviste déclare :

> #Wikibase est une solution évidente pour tout futur logiciel #LOD de description/gestion d'#archives. Un de ces principaux avantages (parmi tant d'autres) est de pouvoir clairement sourcer les déclarations. Les autres : interopérabilité, outils liés, etc. (@B2C, 2019)

Tandis qu'un autre utilisateur – Paul-Olivier Dehaye, fondateur de l'instance Wikibase PersonalData.io – renchérit :

> L'écosystème d'outils qu'il est possible de constituer au sein de Wikibase permet d'aller bien bien au delà du cadre strict des archives! (@podehaye, 2019)

Enfin, Wikibase est également évoqué lors d'échanges portant sur le rôle que pourrait jouer Wikidata dans le cadre du Records Management (@contemplatingIM, 2018), de quoi ouvrir encore davantage les champs d'expérimentations au cours des années à venir.

2.2 Potentiel de Wikibase pour les GLAMs

2.2.1 Maintenance de l'infrastructure

Comme l'ont montré les pages précédentes, dédiées à l'émergence d'une communauté d'utilisateurs Wikibase, les *Galleries, Libraries, Archives, and Museums* (GLAMs) y occupent une place importante. Cette section vise à aborder les possibilités et limites de Wikibase pour ces institutions, en mettant l'accent sur la dimension de maintenance. Plutôt qu'une présentation exhaustive du fonctionnement de Wikibase – des tutoriels existent pour cela –, nous nous concentrons ici sur la question de la maintenance, en l'abordant à travers deux niveaux : du point de vue de l'infrastructure dans la présente sous-section, du point point de vue des données elles-mêmes dans la seconde sous-section.

Alors qu'une vidéo postée sur YouTube illustre comment le logiciel Wikibase peut littéralement être installé en deux minutes [111], Sam Wilson, un ingénieur logiciel de la Fondation Wikimedia, questionne [112] : « Is Wikibase installation *easy* ? » (MediaWiki, 2019). Adam Shorland, développeur de Wikibase employé par Wikimedia Deutschland, lui répond :

> *Installing* is easy, advanced configuration is more advanced, if you want it to be exactly like Wikidata, expect more pain. (MediaWiki, 2019)

Un constat similaire est dressé dans le cadre du projet *Passage* :

> Although the Wikibase editing environment was intuitive and easy to use [113], the pilot participants concluded that it is not ready to serve as turnkey solution for the metadata creation workflow. (Godby *et al.*, 2019, p. 59)

Un commentaire posté en octobre 2018 sur une page de discussion Wikimedia a retenu notre attention. Il a été écrit par Laura Hale, qui a œuvré à la création de l'instance Wikibase *ParaSports* [114] :

> [...] the underlying Wikibase software is not very easy to use. It has a lot of massive fails. The data compression is not greater. The documentation can be awful or non-existent. So much of the software is proprietary that unless you're doing a standalone Wikibase install designed for human reading, it can be impossible without substantial developer investment time and money wise to get it to work. [...] you'd have to draw from people interested in Wikibase generally (of which there are few), people who are in your subject domain with developer skills (of which there may be few) or pay people (who could in theory then be poached by a chapter or the WMF.). The only reason I use Wikibase is out of pure desperation, and its limitations are huge in terms of what we can do without a serious investment of time and money.

111. https://youtu.be/P174BEDhUJg.

112. Cette question est publiée sur une page discussion liée au tableau récapitulatif *Managing data in MediaWiki* (https://www.mediawiki.org/wiki/Manual:Managing_data_in_MediaWiki), dont une ligne, aujourd'hui supprimée, visait à décrire le niveau de complexité d'installation du logiciel. Ce niveau avait été décrit au départ comme *medium* avant de faire l'objet d'une modification et devenir *easy (standard MediaWiki extension install)*, voir : https://www.mediawiki.org/w/index.php?title=Manual:Managing_data_in_MediaWiki&diff=3465012&oldid=3465006.

113. Un questionnaire destiné à comparer les impressions des participants au projet par rapport à d'autres projets de *library linked data* avait par exemple suscité ce commentaire : « Non-buggy, easy to use software » (Godby *et al.*, 2019, p. 58).

114. L'instance n'est plus accessible actuellement, mais un aperçu de l'instance a été archivé : https://web.archive.org/web/20181124014251/https://para-sports.es/wiki/Main_Page.

Ce message est à considérer avec du recul, étant donné qu'il contient une évidente part de subjectivité, qu'il date de 2018, qu'il se rapporte vraisemblablement à une installation réalisée *manuellement* et non pas à l'aide de la version dockerisée [115] et que, comme nous l'avons vu, la communauté Wikibase s'est fortement développée depuis. Il a toutefois le mérite de mettre en avant différents éléménts-clés : des lacunes en matière de documentation ; le caractère *non agnostique* des composants logiciels, conçus spécifiquement pour Wikidata au départ ; l'importance de pouvoir accéder aux connaissances de la communauté Wikibase ; la nécessité de posséder des compétences en développement. Nous nous proposons donc de passer en revue ces différents points et de voir s'ils sont toujours d'actualité.

Tout d'abord, nous rassemblons en un seul et même point la dimension de documentation, de soutien technique et d'entraide au sein de la communauté. En effet, dans les faits, les trois s'interfèrent allègrement. Prenons par exemple le groupe Telegram : il s'avère être à la fois un espace d'échange informel d'expériences et d'astuces entre utilisateurs [116] ; un espace de signalement [117] et d'identification [118] de bugs ; un espace de sollicitation des membres de la communauté [119] ; un espace de réflexion sur les bonnes pratiques [120] et les désidératas de la communauté [121] ; un espace d'assistance

115. Sur cette page datant de fin 2017 et concernant un retour d'expérience basé sur l'installation de l'instance ParaSport Data, nous pouvons lire : « Installation of Wikibase is tricky, a lot of component have been installed to fully use it : MediaWiki, Wikibase Extension, Blazegraph, QueryInterface, Lua ... », voir : https://www.wikidata.org/wiki/Wikidata:WikidataCon_2017/Notes/Wikibase:_How_to_survive_the_install_and_data_normalization_to_get_pretty_research_information.

116. « Hi ! Has anyone, other than Adam, implemented Adam's workaround for resolving URIs in the query service yet ? And have any sorts of difficulties been encountered that a non-programmer (like myself) might be likely to run into when trying to mess around with the docker image ? Thankful for any tips ! » [Message Telegram] Alan, 10.12.19.

117. « Important to everyone running a Wikibase instance with Query Service GUI : we identified a security issue and fixed it. Please update your code ! » [Message Telegram] Léa Auregann, 07.11.19.

118. « Perhaps... I fear we're going to be asking you to open a ticket. You said you are running 1.33 from the tarballs right ? » [Message Telegram] Tom, 20.01.20.

119. « May I cordially invite @all to join the UG by signing the UG page under the *Interested in participating* ? » [Message Telegram] Andra Waagmeester, 20.02.20.

120. « @AndraWaagmeester Other than explicitly mapping Wikidata and satellite-Wikibase identifiers, do you know of other best practices for operating a satellite Wikibase ? » [Message Telegram] James Hare, 04.03.20.

121. « Ok thanks. That's important to know because so far this wasn't on my radar as something people want to do. » [Message Telegram] Lydia Pintscher, 29.10.19.

technique [122] ; un espace de *networking* [123] et d'annonce d'événements [124]. Sans surprise, ce flou suscite de la part des nouveaux venus des questionnements, dont l'éclaircissement repose sur la bonne volonté des membres plus anciens du groupe [125] :

> hello all, where would you like to see reports of bugs and feature requests for Wikibase ? in this group or on https://wikidata.org/wiki/Wikidata:Contact_the_development_team or elsewhere ? [126]

> Is there a helpdesk or the channel to ask questions about for newbie technical problems after a WB docker installation. We could not find in the documentation ? [127]

> [...] Is this the primary communication channel for wikibase tech related stuff ? I've already subscribed the wikidata, wikidata-tech and wikibaseurg mailing lists and am also using #wikidata@freenode. Plus phabricator ;) Where should i head for at first ? [128]

Selon nous, cette confusion provient de deux facteurs : d'une part, la multiplicité des canaux de communication sans que leur spécificité ne soit facilement discernable – a-t-elle seulement été clairement établie ? – pour les nouveaux venus, d'autre part, les lacunes en matière de documentation. Ce problème, déjà souligné en 2017 [129], est toujours présent dans les discussions en 2020 (Wikibase Community, 2020), aux côtés de la question de la pérennité des informations échangées sur le groupe Telegram [130]. Il faut

122. « Hi there, I ask for advice concerning a login problem in QuickStatements : [...] Thanks in advance ! » [Message Telegram] Georg Hertkorn, 20.05.20.

123. « Howdy. Beeing quite inpolite i forgot to introduce me after joining this group : i'm Hans-Jürgen from (near to) Frankfurt/Main, Germany and i'm working at the german national library on [a] Wikibase project that evaluates the use of wikibase for the integrated authority file. Just a hello and nice to meet you :) » [Message Telegram] Hans-Jürgen Becker, 11.05.20.

124. « Tomorrow there will be a small hackathon in the SMW Con in Paris to discuss about LOD, centered about SMW but also with links with the semantic Web and probably Wikibase/Wikidata. » [Message Telegram] Seb35Wikibase, 26.09.19.

125. « This is a good place to get advice in an interactive format, as you have people like Adam who are involved with Wikibase development here. Phabricator is a great place to file a bug, while this is probably the place to find out whether it's a good idea to file a bug, or if there is something else you should be doing to fix X » [Message Telegram] GreenReaper, 11.05.20.

126. [Message Telegram] Rik, 18.02.20.

127. [Message Telegram] JarmoS (sarrit), 20.02.20.

128. [Message Telegram] Hans-Jürgen Becker, 11.05.20.

129. Jan Dittrich WMDE écrit ainsi : « How can I ensure that I checkout the Wikibase version that works for my Mediawiki version ? [...] The documentation is currently not clear on this, but it seems to be essential to get it running » (MediaWiki, 2017).

130. « Among the points discussed, there was a request to log this chat channel or document important issues discussed. Are there idea's on how to this ? We could maybe extract excerpts

savoir que ce problème n'est pas nouveau dans l'univers Wikimedia. Ainsi en 2015, Sebastiaan ter Burg – ancien coordinateur Wikimedia-GLAM aux Pays-Bas – soulignait dans une présentation donnée à destination des GLAMs « The problem within the Wikimedia Projects is WTFM [131] : Write The Fucking Manual. » (Wikimedia Nederland, 2015). En ce qui concerne spécifiquement Wikibase, Adam Shorland – l'un de ses développeurs –, résume ainsi la situation :

> I'd like to think most of the knowledge is out there documented, but it still really isn't pulled together in a good way, there isn't really any focus on that part. [132]

Il faut admettre qu'il est particulièrement bien placé pour le savoir, étant donné que certains points-clés de la documentation [133] sont expliqués directement sur son propre blog de développeur [134]. Ce blog constitue donc l'une des nombreuses ressources détaillant l'une ou l'autre étape de l'installation et de la configuration d'une instance Wikibase… L'Annexe 1 [135] propose un inventaire aussi exhaustif que possible de toutes ces sources éparses – qui s'avèrent parfois incorrectes [136] ou obsolètes [137] –, mêlant une vingtaine de pages d'aide, de tutoriels partagés par des utilisateurs sous forme de blog ou de vlog, de discussions de groupe, ou encore d'instructions données sur un dépôt GitHub.

Les choses évoluent toutefois : l'équipe derrière Wikibase est consciente du problème [138], elle a mis en place une refonte du site web officiel de Wikibase afin d'aider les utilisateurs à trouver les outils et la documentation nécessaires [139], et enfin, un message posté le 19 mai 2020 sur la mailing Wikibase [140] annonce l'arrivée d'un nouveau rédacteur technique pour Wiki-

from this channel and store them in a separate part of the UG Wiki ? » [Message Telegram] Andra Waagmeester, 20.02.20.

131. Un détournement du sigle anglais RTFM, pour *Read The Fucking Manual*.

132. [Message Telegram] Adam, 07.09.19.

133. Comme par exemple : *Changing the concept URI of an existing Wikibase with data*.

134. Voir : https://addshore.com/2019/11/changing-the-concept-uri-of-an-existing-wikibase-with-data/.

135. Page 321.

136. L'utilisateur Telegram Myst témoigne par exemple avoir dû corriger des exemples qui n'étaient pas fonctionnels : https://github.com/samu-workopen/learningwikibase/pull/21/files.

137. C'est le cas par exemple du fichier d'installation Docker qui ne pointait pas vers la version la plus à jour : https://github.com/wmde/wikibase-docker/commit/bcd08f9b2fb4b758ab8aeba1978fb016297ce004#diff-4e5e90c6228fd48698d074241c2ba760.

138. « We don't have a point where people go to after they've installed Wikibase to see the most common next steps. We should have that. » https://phabricator.wikimedia.org/T231191.

139. « The Wikibase website needs an overhaul to be a better entry to the world of Wikibase and help people find all the tools and documentation around it. » https://phabricator.wikimedia.org/T205605.

140. https://lists.wikimedia.org/pipermail/wikibaseug/2020-May/000077.html.

data et Wikibase, dont le premier objectif sera de collecter et d'améliorer la documentation concernant les étapes de post-installation de Wikibase.

Deuxièmement, outre la documentation, l'un des autres obstacles à l'utilisation de Wikibase est le savoir-faire technique requis. Certes, l'utilisation d'une Wikibase est possible sans connaissances particulières liées au Web sémantique, comme le souligne le rapport du projet Passage :

> Wikibase also offers many technical advantages as a platform for experimentation with library resource-description workflows [...] Most important, this technical detail is mostly hidden from human users of the Wikibase applications. Thus, metadata librarians familiar with current workflows can easily interact with the editing interface to create resource descriptions in a new idiom. In other words, the Wikibase platform offers the promise that the transformation from human-readable records to machine-understandable knowledge graphs can occur in the library metadata creation workflow without presupposing any knowledge of RDF, Turtle, triples, or other details in a linked data implementation. (Godby *et al.*, 2019, p. 12)

Cependant, ces détails techniques *majoritairement cachés* doivent toutefois être pris en charge et l'instance Wikibase doit être configurée. Alors que le site officiel Wikibase fait miroiter une installation *facile* pour attirer le chaland [141], Allison-Cassin et Seeman nuancent : « it's easy in that it's possible but results may vary depending on your technical expertise and background » (Allison-Cassin et Seeman, 2019). C'est également ce qui ressort de cette intervention sur le groupe Telegram Wikibase, portant sur les obstacles rencontrés par le créateur de l'instance PersonalData.io :

> Get just a basic running instance going. That is painfully hard. You need to get a good dev somehow to fight with the Docker image, etc. And it is really really really stupid that it should be so hard. [142]

Or, bien que l'utilisation de Docker présente déjà son lot de difficultés [143], cela ne s'arrête pas là :

141. « Use our quick and easy setup with Docker », peut-on lire sur https://wikiba.se/faq/.
142. [Message Telegram] Paul-Olivier Dehaye, 07.09.19.
143. Un développeur œuvrant à la maintenance d'une instance Wikibase, qui n'avait pas d'expérience préalable avec Docker, nous a par exemple confié lors d'une visioconférence [réalisée via Telegram, le 26 mars 2020] que s'il trouvait Wikibase relativement stable pour les utilisateurs, il avait en revanche constaté que du point de vue de l'administrateur système, des erreurs pouvaient être très vite commises (comme par exemple tout supprimer à cause d'une commande hasardeuse). Il soulignait dès lors la nécessité d'effectuer des sauvegardes régulières de toute l'installation.

> The Docker part is fine. It's relatively easy to find devs who know Docker. The problem is the idiosyncrasies of MediaWiki - Wikibase, combined with sometimes missing sometimes incorrect documentation. [144]

Or, se familiariser avec MediaWiki requiert des efforts, comme le souligne l'utilisateur GreenReaper :

> Unfortunately MediaWiki itself has a pretty big learning curve, so you end up having to learn two complex things when using Wikibase. [...] a lot of things are setup for pure MediaWiki and the implications of Wikibase being in the middle have only been fully explored by Wikidata (which is full of people who « just know » the answer). [145]

De plus, il faut savoir que Wikibase fonctionne en réalité à l'aide de l'enchevêtrement de deux extensions, ce qui génère encore davantage de complexité [146]. La structure de ces extensions est d'ailleurs questionnée par l'équipe de développement elle-même [147]. Cet entremêlement à MediaWiki impose également aux administrateurs d'instances Wikibase de se familiariser avec de nouveaux usages, comme par exemple le fait d'investir la plateforme *Phabricator* pour consulter des rapports de bugs, signaler un problème ou suggérer de nouvelles fonctionnalités.

Par ailleurs, le fait que Wikibase soit un logiciel libre et non un produit commercial, signifie qu'il n'existe pas d'assistance technique disponible sur demande, à part peut-être dans le cadre de partenariats officiels entre institutions et Wikimedia [148]. Alors certes, une certaine entraide – assez caractéristique des dynamiques à l'œuvre au sein des communautés du logiciel libre (Demazière *et al.*, 2006) – prend place entre utilisateurs, et les développeurs sont présents, tant sur la plateforme Phabricator que sur le groupe Telegram

144. [Message Telegram] Paul-Olivier Dehaye, 07.09.19.
145. [Message Telegram] GreenReaper, 27.01.20.
146. « L'utilisation du Wikibase Client demande obligatoirement un partage de la base de données du Wikibase Repository. L'installation de cette extension est dans tous les cas compliquée et très mal documentée. Elle demande aussi de paramétrer son installation Mediawiki de façon particulière, comme l'activation des interwikis (même si il n'y a qu'une instance). Encore quelque chose tourné principalement vers Wikidata et non Wikibase. » [Message Telegram, échange privé] Myst, 26.05.20.
147. « Currently Wikibase extension is conceptually divided into Repo and Client components. These components are not clearly separated, interdependent, and often, intentionally and not intentionally use the same code pieces. This entanglement affects negatively the productivity when making change to the Client part (might be non intentionally affecting *Repo*), and the other way round. » (Wikidata, 2020b).
148. Ainsi, sur cette page de l'instance FactGrid inventoriant problèmes et désidératas, nous retrouvons plusieurs réponses et commentaires émis par des membres de l'équipe de développement de Wikibase, voir : https://database.factgrid.de/wiki/FactGrid:Troubleshooting.

ou dans des réponses à la *mailing list*, mais cela reste parfois ambigu [149] et surtout aléatoire [150]. C'est d'ailleurs ce constat qui a conduit à la création – bénévole – de la plateforme de documentation LearningWikibase. Son initiatrice, Sandra Müllrick, relève sur le dépôt GitHub associé à la plateforme (Müllrick, 2019) :

— Technical interaction with the Wikibase Software is challenging
— Wikibase is used very individual. Users have lots of different requirements, so there are a lot of different use cases
— Wikibase user can't find enough developer support to get help in setting up a Wikibase ins[t]ance.

C'est précisément le phénomène résultant de cela qui a justifié la création de cette plateforme de documentation, à savoir le fait qu'un petit groupe d'utilisateurs pionniers de Wikibase se retrouvaient énormément sollicités par de nouveaux utilisateurs ayant besoin de guidance personnalisée, la documentation existante ne leur suffisant pas (Müllrick, 2019).

Enfin, outre la création et maintenance quotidienne d'une instance Wikibase, avec toute la gestion des bugs et mises à jour que cela sous-entend (Aeyers *et al.*, 2019; Shorland, 2019b), ainsi que les aspects liés spécifiquement à des installations *annexes* qui seront abordés au cours des prochains paragraphes, il faut prendre en considération le fait que les fonctionnalités disponibles ne sont pas toujours suffisantes. Si elles ne comblent pas complètement les besoins de l'organisation utilisant l'instance, cette dernière pourrait souhaiter que des solutions *ad hoc* soient testées. C'est le cas par exemple de l'ABES et de la BnF, qui ont constaté « [a] gap between the specific needs of our institutions and the solutions offered by Wikibase » [151] (Angjeli et Bober, 2019, p. 4).

149. Ainsi, l'un des développeurs de Wikimedia Deutschland affiche sur son blog un lien vers sa page *Buy me a coffee* – cette plateforme permet aux internautes d'apporter leur soutien à des créateurs au sens large – : il apparaît qu'un utilisateur a souhaité le remercier pour son soutien : « Thank you very much for setting up WBStack in general, and helping me set up my wiki ». Ce qui entraîne la question : est-ce qu'apporter une assistance aux utilisateurs fait partie de son cahier des charges ou le fait-il de façon bénévole en marge de ses autres tâches ?

150. Ainsi, cette demande envoyée sur la *mailing list* en octobre 2019 n'a par exemple jamais fait l'objet d'une réponse – publique du moins. Voir : https://lists.wikimedia.org/pipermail/wikibaseug/2019-October/000045.html.

151. Ainsi, en ce qui concerne la question des *rôles et restrictions dans Wikibase*, ABES et BnF avaient notamment précisé aux développeurs associés au projet que certaines données ne devaient être visibles que par une partie des utilisateurs – à savoir, les utilisateurs *logués* seulement (ABES et BNF, 2019b). Ce qui avait amené l'un des développeurs à conclure « Aucune solution préexistante ne semble être disponible pour couvrir les besoins décrits. La sécurisation des données sensibles n'est pas dans les objectifs de Wikibase et nous savions que le poc [proof of concept] allait coincer sur ce sujet. » (ABES et BNF, 2019b).

Or, les pistes envisagées pour faire face à de telles situations se basent sur des ajustements [152] qui nécessitent des compétences et ressources que toutes les institutions sont loin de posséder.

Avant de clôturer ces paragaphes dédiés aux compétences techniques qu'il semble inévitable de posséder pour pouvoir se lancer dans la gestion d'une Wikibase, il est toutefois nécessaire d'attirer l'attention sur WBStack [153], qui a été brièvement évoqué au cours des pages précédentes. Ce service lancé en octobre 2019 vise précisément à délester les utilisateurs potentiels de Wikibase de toute contrainte technique :

> The idea behind the project is to provide Wikibase and surrounding services, such as a blazegraph query service, query service ui, quick statements, and others on a shared platform where installs, upgrades and maintenance are handeled centrally. (Shorland, 2019a)

La plateforme, qui a déjà permis la création de plus de 200 instances (Shorland, 2020b), est sponsorisée par l'organisation Rhizome et développée par Adam Shorland, développeur Wikimedia Deuschland qui explique travailler sur ce projet pendant son temps libre (Shorland, 2019a). Cela pose bien entendu la question de sa pérennité et de sa gratuité : que se passerait-il en cas de désinvestissement de ce développeur ? Quelles sont les capacités techniques de la plateforme pour faire face à une demande potentiellement croissante ? Quelles assurances pour les institutions qui privilégieraient cette option, en matière de stabilité, de performance, de sauvegarde, de gratuité et de pérennité ? Contactée à ce sujet en mai 2020, Lydia Pintscher – *Wikidata product manager* – explique qu'aucune décision n'a encore été prise à ce sujet [154]. La fin de l'année 2020 devrait toutefois permettre de clarifier ces questions : « WMDE work around *Wikibase as a Service* is planned in this area during the second half of 2020 » (Shorland, 2020a).

Enfin, une troisième limite liée à l'implémentation de Wikibase est ce que nous pourrions désigner comme le manque de maturité du logiciel, ou plutôt son manque d'adéquation à des usages hors de la sphère de Wiki-

152. Dans le cas précédemment évoqué, l'une des pistes évoquées était « un proxy qui viendrait filtrer les requêtes sur certaines pages, en dehors de Wikibase pour éviter les contournements internes découragés » (ABES et BNF, 2019b), mais nous pouvons également penser au système conçu pour la Wikibase *WikiLex* (https://git.en-root.org/Seb35/wikilex-sync) qui permet la synchronisation entre des bases de données préexistante et l'instance Wikibase, ou encore songer aux développements annexes réalisés dans le cadre du projet Passage : « During the Passage pilot, [...] the Wikibase software suite was extended with the Explorer and Retriever, two new applications developed at OCLC that fill in gaps in the library resource-description workflow » (Godby *et al.*, 2019, p. 18).

153. https://www.wbstack.com/.

154. « We are currently looking at all options but have not made a decision yet. » [Message privé Telegram] Lydia Pintscher, 19.05.20.

data. Comme souligné au fil des pages de la section précédente, Wikibase a été conçu expressément pour répondre aux besoins du projet Wikidata. Cela signifie que plusieurs éléments constitutifs du logiciel ou des applications annexes sont encore marqués par ce caractère « sur mesure », requérant dès lors des ressources supplémentaires pour être adaptés à une instance Wikibase indépendante. C'est sans doute ce qui a amené – comme nous l'avons vu quelques paragraphes plus tôt – l'utilisatrice Laura Hale à déplorer que « so much of the software is proprietary », alors même que tout le code issu des projets Wikimedia est pourtant libre et *open source* [155].

Ce caractère *non agnostique* de Wikibase se manifeste à différents niveaux. Premièrement, les pages d'aide associées à une instance Wikibase sont par défaut des renvois vers des ressources produites par la communauté Wikidata [156] et ne sont dès lors pas forcément adaptées à un contexte de Wikibase indépendante. Le projet Passage explique ainsi avoir dû modifier certaines ressources issues de Wikidata, comme le portail d'aide ou le glossaire (Godby *et al.*, 2019, p. 18). C'est le cas également d'éléments peut-être plus anecdotiques, tels que les références à Wikidata par défaut pour des éléments concernant des dates, des quantités ou des coordonnées géographiques [157].

Deuxièmement, le *SPARQL endpoint* compris dans la version *dockerisée* de la suite logicielle Wikibase possède lui aussi des éléments propres à Wikidata. Comme le relève l'équipe de développement qui a inscrit cet objectif dans son plan de travail pour 2020 :

> Make Query Service less specific to Wikidata : to better suit the needs of Wikibase use cases outside of Wikidata.org, we will make changes to the Query Service to move defaults away from Wikidata and toward Wikibase. (Wikidata, 2020b)

Concrètement, cela se traduit par le fait que la documentation et les exemples donnés soient spécifiques à Wikidata [158] et que les préfixes par défaut soient également ceux de Wikidata [159]. Par ailleurs, d'autres difficultés propres à

155. https://www.mediawiki.org/wiki/How_to_contribute.

156. Comme par exemple https://www.mediawiki.org/wiki/Help:Contents

157. Par exemple, dans cet export en format JSON de l'élément FactGrid Q105760 (https://database.factgrid.de/wiki/Special:EntityData/Q105760.ttl), il apparaît que les informations additionnelles permettant de préciser la nature des coordonnées géographiques (voir https://en.wikibooks.org/wiki/SPARQL/WIKIDATA_Precision,_Units_and_Coordinates#Coordinates font référence à un globe – *wikibase :geoGlobe* – et que la valeur par défaut de ce dernier n'est autre que l'élément Q2|terre, directement issu de Wikidata (http://www.wikidata.org/entity/Q2).

158. Voir par exemple cette tâche Phabricator encore non résolue : « configure Factgrid Query Service UI to use local example queries », voir : https://phabricator.wikimedia.org/T223586.

159. « If it's your own Wikibase, *wd :* and *wdt :* won't work, unless you're made a different PREFIX declaration in your query. [...] Part of why this is somewhat mysterious is that in

Wikibase ont été signalées, comme par exemple ce problème dans l'affichage de résultats contenant des dates (en fonction de leur degré de précision) (FactGrid, 2020a, p. 4).

Troisièmement, l'outil *QuickStatements*, installé par défaut dans le cadre de la version *dockerisée* de Wikibase et destiné à l'édition massive d'éléments Wikibase, présente différentes difficultés : il a été longtemps non opérationnel [160], il est source d'erreurs pour des opérations fonctionnant correctement dans le cadre de Wikidata [161] et enfin, son interface n'a pas été adaptée à un usage hors de Wikidata : outre un descriptif faisant explicitement référence à Wikidata (« QuickStatements is a tool to batch-edit Wikidata »), certains boutons – comme *dernier lots* – ne sont pas fonctionnels par défaut et doivent vraisemblablement faire l'objet de modifications [162].

Quatrièmement, comme souligné dans le plan de développement de Wikidata et Wikibase pour l'année 2020, « Wikibase users, particularly those in the GLAM sector, want to be able to link to/display media in Wikibase that is not and cannot be on Commons » (Wikidata, 2020b). Or, à l'heure actuelle, il est n'est pas possible d'*uploader* et d'afficher ses propres images sans passer par Wikimedia Commons [163]. Outre les images, l'interface graphique Wikibase pose d'autres problèmes d'affichage : par exemple, les coordonnées géographiques sont accompagnées de leur représentation sur une carte sur Wikidata [164], mais cela n'est pas implémenté par défaut sur les instances Wikibase (FactGrid, 2020a, p. 1).

Cinquièmement, il existe un certain nombre d'extensions, de gadgets et d'applications développés pour améliorer l'expérience utilisateurs sur Wikidata ou pour déployer de nouveaux outils associés aux données contenues dans la base de connaissance. Si leur réutilisation est théoriquement possible et figure parmi les arguments en faveur d'une utilisation de Wikibase par les GLAMs (Fischer et Ohlig, 2019), ils ne sont en revanche pas forcément prêts à l'emploi et susceptibles de nécessiter des ajustements. Comme le souligne Fischer :

> We at the German National Library had imagined many things to be easier. Precisely because Wikidata itself is a very powerful

Wikidata, we benefit from the many prefixes that are implicit/pre-loaded without exposing it to the user. Wd, wdt, p, ps, pq, et al. So when you move to your own Wikibase, you have to know a lot more about this » [Message Telegram] Andrew Lih, 23.11.19.

160. En raison de paramètres liées à OAuth.

161. Voir par exemple : https://phabricator.wikimedia.org/T250143.

162. https://database.factgrid.de/wiki/FactGrid:Setup#QuickStatements.

163. Comme en témoignent par exemple ces deux tâches Phabricator encore en cours de traitement https://phabricator.wikimedia.org/T90492 et https://phabricator.wikimedia.org/T251021

164. Voir par exemple cette carte https://www.wikidata.org/wiki/Q573179#/map/0 issue des coordonnées géographiques liées à l'élément Q573179|Couillet (https://www.wikidata.org/wiki/Q573179).

> database, we would have thought that importing large amounts of data was already part of the standard implementation. Perhaps it was our perspective as a cultural institution and our lack of experience with an open application such as Wikibase with its close integration with the Mediawiki software that initially seemed unusual. Such a close connection, also inscribed in the code, is normal for Wikipedia and many other wikis, but it is unexpected when one reckons with database software. The development is directly driven by a voluntary community, which also explains why many of the attractive additions are still Wikidata-affine and would first have to be adapted for generic re-use by all Wikibase users. (Fischer et Ohlig, 2020)

Si pour certains de ces outils, un *simple* ajout du gadget ou de l'extension suffit [165], pour d'autres, quelques éléments du code doivent être préalablement adaptés [166]. Enfin, pour d'autres encore, les modifications nécessaires sont telles que l'outil ne peut pour l'instant pas du tout être utilisé [167].

Finalement, bien que le tableau puisse paraître particulièrement sombre à la lecture de ses différents obstacles, il ne faut toutefois pas perdre de vue que le lancement de la première version *dockerisée* de Wikibase remonte à octobre 2017 seulement, que l'équipe de développement se montre à l'écoute des besoins des utilisateurs [168], et surtout déterminée à investir dans le perfectionnement du logiciel avant que d'autres acteurs commerciaux ne lui ravissent la place (Pintscher *et al.*, 2019b, p. 8).

2.2.2 Maintenance des données

Le fait de bénéficier d'une instance Wikibase vierge, indépendante de Wikidata, signifie ne pas être soumis aux critères de notoriété [169] établis par

165. C'est le cas par exemple du gadget *Description* nécessitant seulement que le code Javascript soit copié sur une page dédiée, voir par exemple : https://wiki.personaldata.io/wiki/MediaWiki:Gadget-Descriptions.js

166. À l'instar du gadget *Easy-Query*, voir https://wiki.personaldata.io/wiki/Wikibase_options#Gadgets ou de l'application *Reasonator* (FactGrid, 2020a, p. 5)

167. C'est le cas par exemple du logiciel OpenRefine qui possède une extension Wikidata permettant de transformer des données tabulaires en modifications de Wikidata, mais dont l'adaptation à d'autres instances Wikibase est loin d'être triviale, comme l'explique le développeur Antonin Delpeuch : « it [would be] a significant amount of work to do it properly » (Antonin, 2018). Mise à jour septembre 2020 : cette situation a évolué depuis et l'adaptation a pu être réalisée.

168. « Today on the Wikibase Community User Group Telegram chat I noticed some people discussing issues with upgrading Mediawiki and Wikibase using the docker images provided for Wikibase », explique par exemple le développeur Adam Shorland (Shorland, 2019b)

169. Ces derniers servent à déterminer si une donnée a sa place sur Wikidata. Un élément est jugé acceptable s'il contribue à l'un des deux objectifs principaux de Wikidata (centraliser les liens interlangues à travers les projets Wikimedia et servir de base générale de connaissance

la communauté Wikidata. Créer sa propre instance Wikibase apparaît donc comme une alternative aux yeux des personnes et organismes dont les données ne répondent pas toujours à ces critères, à l'instar du projet Linked Jazz [170]. De plus, le fait de ne pas être astreint à respecter ces critères peut permettre de dépasser certains problèmes d'inclusion et de diversité ayant été dénoncés, notamment en ce qui concerne les données liées à des personnes (Allison-Cassin *et al.*, 2019, p. 11). Par ailleurs, outre ses critères de notoriété, il ne faut pas oublier que Wikidata se caractérise avant tout par le fait de ne publier que des données publiées sous une licence Creative Commons CC0 [171], c'est-à-dire, que les données sont transférées dans le domaine public [172]. Or, si certaines institutions adoptent des pratiques similaires, à l'instar de la Bibliothèque nationale allemande [173], il est clair que toutes ne sont pas prêtes ou en mesure d'adopter une politique si tranchée en matière de droits d'auteur [174] (Waagmeester *et al.*, 2018b).

Le fait de bénéficier d'une autonomie totale quand à la publication des données entraîne dès lors des questions liées à leur contrôle et à leur maintenance. Il serait fastidieux d'analyser en détail comment chaque instance Wikibase gère l'intégralité des droits de ses utilisateurs [175]. En revanche, il semble intéressant de se pencher sur la question de l'édition des données. En effet, Wikibase est présenté comme un outil permettant de déployer une base de connaissance collaborative et ses atouts en matière de crowdsourcing ont été soulignés à plusieurs reprises, notamment dans le contexte de la gestion des données d'autorité :

---

pour tout le monde) et répond dès lors au minimum à l'un de ces trois critères : il contient au moins un lien de site valide vers une page des projets Wikimedia ; il fait référence à une instance d'une entité matérielle ou conceptuelle clairement identifiable ; il remplit un besoin structurel, par exemple lorsqu'il est nécessaire pour rendre plus utiles des déclarations faites dans d'autres éléments. (Wikidata, 2020e).

170. Miller explique ainsi « for us, we are going to have a lot of esoteric data, for example modeling oral history transcripts down to the statement level. We will end up storing a lot of data that we use to power our tools and research but is really not appropriate to put into Wikidata. » (Miller, 2018).

171. Voir : https://www.wikidata.org/wiki/Wikidata:Copyright.

172. Pintscher, *Product Manager for Wikidata*, en résume ainsi la raison : « Wikidata is here to give more people more access to more knowledge. This means we want our data to be used as widely as possible. CC-0 is one step towards that. » (Pintscher, 2017).

173. Dont le GND est d'ores et déjà publié sous une licence *Creative Commons Zero*, voir : https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.

174. À titre illustratif, nous pouvons par exemple citer l'instance Biblissima, qui précise sur sa page d'accueil (https://data.biblissima.fr/w/Accueil) que tous les référentiels sont publiés selon les termes de la Licence Ouverte version 2.0 (Etalab), qui se distingue de la licence CC0 par une « exigence forte de transparence de la donnée et de qualité des sources », « rendant obligatoire la mention de la paternité » (voir : https://www.etalab.gouv.fr/licence-ouverte-open-licence).

175. Ces derniers peuvent être consultés en complétant l'URL de chaque Wikibase à l'aide de : wiki/Special:ListGroupRights.

> The Wikibase implementation is attractive because it reflects a sophisticated understanding of what is required to support crowdsourcing across a global community of users, which is being evaluated by OCLC and many libraries as a model for modernizing the practice of cooperative cataloging and authority management. (Godby *et al.*, 2019, p.8)

Il ne s'agit dès lors pas seulement d'une question technique, mais également d'une question de politique éditoriale. Il s'agit de déterminer qui peut créer ou modifier des données, et à quelles conditions. Dans les faits, il s'avère qu'il y a tout un spectre de possibilités, de la solution la plus ouverte et permissive aux choix les plus fermés et restrictifs.

À l'extrémité la plus latitudinaire du spectre, se situe Wikidata, cette « base de connaissances libre et gratuite qui peut être lue et modifiée tant par les personnes que par les dispositifs informatisés » [176]. Et par « personnes », il faut entendre : tout le monde, c'est-à-dire toute personne désireuse de contribuer, sans que ne soit requise ni inscription ni compétence particulière [177], bien que ces principes soient régulièrement remis en question [178] et qu'il faille par ailleurs noter que la base de connaissance applique depuis 2014 une *politique de protection des pages* [179]. Malgré cela, Wikidata demeure toutefois l'exemple le plus iconique d'une instance Wikibase favorisant la participation, en étant généreuse dans sa gestion des droits des utilisateurs et en limitant les restrictions.

Ce n'est évidemment pas le cas de toutes les instances Wikibase ayant vu le jour, qui sont libres de gérer ces droits à leur guise. Si nous reprenions cette idée d'un spectre de possibilités, nous trouverions ainsi au centre des instances proposant des formes d'édition collaborative plus *modérées*. Leur démarche consiste à commencer par *filtrer* les contributions en offrant seulement des droits de lecture aux utilisateurs non inscrits, qui devront commencer par se créer un compte et se connecter s'ils souhaitent contribuer. Au sein même de cette approche, plusieurs nuances sont possibles. Par exemple,

176. https://www.wikidata.org/wiki/Wikidata:Main_Page?uselang=fr.
177. https://www.wikidata.org/wiki/Wikidata:Contribute/fr.
178. Ainsi, dans le cadre de Wikipédia, le débat refait régulièrement surface (au point d'être considéré comme une *discussion marronier* (Wikipédia, 2020a)) : une personne ne possédant pas encore de compte utilisateur devrait-elle être autorisée à créer ou modifier des pages Wikipédia ? Actuellement, et conformément au troisième principe fondateur – the ability for almost anyone to edit (most) articles without registration – des projets des Wikimedia (Wikimedia, 2020), il est possible de contribuer sans posséder de compte utilisateur : c'est alors l'adresse IP qui sert d'identifiant par défaut. L'alternative défendue par certains consisterait à interdire la rédaction aux utilisateurs ne disposant pas encore de compte (Wikipédia, 2020b). Si nous n'avons pas déniché de synthèse résumant ce débat dans le cadre de Wikidata, les considérations semblent toutefois similaires (Wikidata, 2020).
179. Voir : https://www.wikidata.org/wiki/Wikidata:Page_protection_policy/fr.

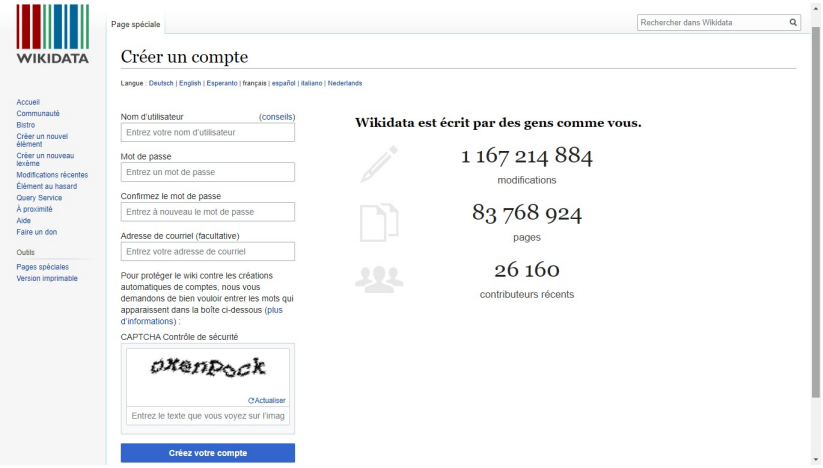


FIGURE 2.10 – Pour lutter contre la création de comptes indésirables, Wikidata recourt à un système de validation Captcha. Source : Wikidata (https://www.wikidata.org).

l'instance Plantdata.io [180] limite les possibilités d'édition de données aux utilisateurs enregistrés. La création de compte nécessite uniquement de choisir un nom d'utilisateur et un mot de passe... Ce qui semble ouvrir la porte aux spams [181]. Personadata.io, qui faisait face au même problème [182], utilise désormais une extension [183], qui « rend obligatoire la demande et l'approbation des comptes ». Une alternative pour lutter contre les spams consiste à installer un système de captcha [184], à l'instar de ce que propose Wikidata [185], comme le montre la figure 2.10.

Comme Personaldata.io, Wikidocumentaries restreint les possibilités de contribution aux utilisateurs dotés d'un compte personnel, mais cette fois-ci le formulaire est plus conséquent [186]. Outre un nom d'utilisateur et une adresse email, il faut également renseigner d'autres informations personnelles, comme un nom officiel, une biographie personnelle, éventuellement un CV et des notes additionnelles et enfin, la confirmation que les conditions d'utilisation sont acceptées... alors même que celles-ci sont, pour l'heure, in-

180. Qui ne semble plus active et dont la page d'accueil n'a pas été modifiée depuis 2017, voir : http://wikibase.plantdata.io/wiki/Main_Page.
181. En décembre 2019 nous avons ainsi constaté la présence d'une page aujourd'hui supprimée, contenant uniquement un lien intitulé *free casino games onlines*.
182. Comme le montre par l'exemple cette archive de page d'un compte utilisateur qui a depuis été bloqué : https://wiki.personaldata.io/wiki/User:EuniceGarretson.
183. ConfirmAccount, voir : https://www.mediawiki.org/wiki/Extension:ConfirmAccount.
184. Par exemple à l'aide de l'extension *ConfirmationEdit*, voir : https://www.mediawiki.org/wiki/Extension:ConfirmEdit.
185. https://www.wikidata.org/wiki/Special:Captcha/help.
186. http://wikidocumentaries.wmflabs.org/wiki/Special:RequestAccount.

existantes [187]. En revanche, du côté de FactGrid, les conditions d'utilisation sont clairement précisées :

> Please be prepared to use FactGrid transparently and as a colleague. We would like to use your real name, offer an address on which others can contact you, handle your projects with openness of FactGrid. (FactGrid, 2020b)

Elles permettent notamment de mieux comprendre pourquoi le droit à la modification des données est restreint :

> FactGrid is primarily a research site. [...] We explicitly encourage « original research », research you should be able to safely publish here for the first time. This is basically possible under the decision to restrict participation to people who will personally sign responsible for what they do on FactGrid, privately or as part of their research project. (FactGrid, 2020b)

Si poser un tel cadre semble porteur [188], cela ne correspond toutefois pas à la politique de toutes les bases de connaissance propulsées par Wikibase. Ce qui nous amène à l'autre bout du spectre, où les droits d'édition sont beaucoup plus verrouillés. C'est le cas par exemple de l'instance utilisée dans le cadre du Leibniz's Correspondents and Acquaintances (LCA) project [189], qui a désactivé la fonctionnalité de création de compte installée par défaut [190] ou encore de l'instance The EU Knowledge Graph, qui invite toutefois les personnes intéressées à contribuer à se manifester [191].

Outre ce contrôle pouvant être effectué en choisissant qui a le droit d'éditer les données ou non, d'autres possibilités existent. L'environnement MediaWiki propose en effet toute une série de rôles [192] auxquels peuvent être accordées diverses permissions, comme par exemple le fait de créer, modifier, fusionner ou supprimer des éléments [193]. Or, la gestion des droits des utilisateurs fait partie des éléments les plus cruciaux si nous pensons aux institutions culturelles investies dans la description de leurs collections. Elles

187. Le formulaire renvoie vers une page type n'ayant pas encore été remplie : http://wikidocumentaries.wmflabs.org/wiki/Project:Terms_of_Service.
188. La page d'accueil de l'instance indique qu'en avril 2020, plus de 100 utilisateurs ont généré plus de 150000 éléments : https://database.factgrid.de.
189. https://leibnitiana.eu/about.
190. La dimension d'édition collaborative ne semble pas à l'ordre du jour : « Each LCA file has an author and data curator, who can consult, add and modify in the Wikibase » (Leibnitiana.eu, 2019).
191. « New knowledge will be added continuously. If you are interested to collaborate, please contact us [...] », voir : https://linkedopendata.eu/wiki/The_EU_Knowledge_Graph.
192. Par défaut : utilisateurs non enregistrés, utilisateurs autoconfirmés ; utilisateurs enregistrés ; robots ; bureaucrates ; administrateurs.
193. Pour une vue détaillée, voir : https://www.wikidata.org/wiki/Wikidata:User_access_levels/fr.

sont par ailleurs susceptibles de déjà utiliser un système sophistiqué de gestion des droits et accès aux données. C'est le cas par exemple de la Bibliothèque nationale allemande [194], qui se demandait en mai 2019 si Wikibase pouvait garantir la fiabilité du contrôle des données d'autorité des bibliothèques par le biais de son système de rôles et permissions (Fischer et Ohlig, 2019). À l'heure du bilan, en mars 2020, on peut lire :

> The sets of rules that are necessary to create truly binding and reliable authority control data can be transferred to a Wikibase instance. Based on our tests so far, we believe that we could use the rule sets modeled in Wikibase to control intelligent and dynamically adapting input masks. (Fischer et Ohlig, 2020)

Certaines réserves ont toutefois été émises par les représentants d'autres bibliothèques nationales, notamment en ce qui concerne le niveau de granularité des droits d'accès [195]. D'autres points d'attention se rapportent à la possibilité de pouvoir limiter la visibilité d'une partie des métadonnées à certains types d'utilisateurs ou de différer leur visilité – notamment dans le cadre d'archives dont la consultation est soumise à une restriction. Enfin, il a été mis en évidence que :

> Permissions might enable other people *in the organization* to contribute who currently are not. [...] Sometimes permissions are helpful - because they give people the responsibility / job title to contribute. (Aeyers *et al.*, 2019, p. 19)

Ce dernier point montre qu'au-delà de l'utilité de la gestion des droits d'accès et de modification pour la maintenance des données, une utilisation avisée de ces rôles et droits peut constituer un levier de motivation et ainsi permettre d'augmenter le taux d'engagement des utilisateurs d'une Wikibase.

194. Comme en atteste la description des différents niveaux de catalogage associés aux personnes éditant le GND, en fonction de leur formation et responsabilités, voir les pages 2 et 3 de ce document : https://wiki.dnb.de/download/attachments/50759357/005.pdf.

195. « User rights management [is] insufficiently nuanced e.g. different levels of access on a per-property basis », relèvent par exemple Pascal Lefeuvre et Tiphaine Foucher du Département des Métadonnées de la BnF (Aeyers *et al.*, 2019, p. 6).

Deuxième partie

Étude de cas : les données nominatives du CegeSoma

Introduction

Cette seconde partie présente une étude de cas. Elle vise à mettre à l'épreuve les hypothèses présentées dans l'introduction. Plus concrètement, il s'agit de tester à l'aide de données empiriques les possibilités et les limites d'une approche faisant appel à une instance Wikibase pour la gestion des données d'autorité archivistiques, en l'occurence, les données d'autorité du CeGeSoma relatives à des personnes physiques.

Cette seconde partie s'ouvre sur le chapitre 3, qui décrit l'institution et son contexte, mais également ses besoins et les besoins de ses utilisateurs. Le corpus de données, son traitement et sa modélisation sont présentés dans le chapitre 4, tandis que l'implémentation d'une Wikibase pour gérer ces données est présentée au cours du chapitre 5. Enfin, le sixième et dernier chapitre repose sur une analyse approfondie aboutissant à une série de recommandations généralisables à d'autres institutions concernées par ces défis.

En préambule à ces quatre chapitres, la figure 2.11 propose une vue d'ensemble des différents composants et des interactions prenant place dans le cadre de cette instance Wikibase. En effet, si nous reviendrons de manière plus détaillée sur la majorité de ces éléments au cours de cette seconde partie, nous trouvions utile de commencer à en esquisser les contours dès cette introduction.

Concrètement, le haut du schéma présente les trois profils d'utilisateurs susceptibles d'interagir avec les données : le personnel qui est amené à la fois à produire des données et à les consulter ; des chercheurs possédant une expertise sur certaines données pourraient être amenés à les éditer, mais également à les consulter, notamment en utilisant un point d'accès SPARQL ; enfin il y a le grand public, habitué à consulter les données à l'aide du moteur de recherche d'un catalogue en ligne.

Sous ce premier niveau se trouvent, de gauche à droite, l'instance Wikibase combinant une interface utilisateur et un système de stockage, dont les données sont converties en triplets RDF [196] pour alimenter un *triplestore*. À ce dernier est adossé un point d'accès SPARQL permettant de récupérer et de manipuler ces données RDF à l'aide d'un langage sémantique de requête. Ces requêtes SPARQL sont également utilisées par le moteur de recherche des Archives de l'État pour récupérer des données destinées à être affichées – aux côtés des données *locales* issues des fichiers d'indexation des collections – dans son catalogue en ligne, Search. À noter que ces requêtes peuvent

196. Il s'agit d'une conversion en RDF spécifique au modèle de données Wikibase, dont les triplets peuvent être augmentés par des données *annexes* comme des références ou des qualificatifs, ainsi que nous l'avons vu au cours du chapitre 2.

potentiellement être formulées sous forme de requêtes fédérées permettant d'inclure aux résultats des données issues de bases de connaissance externes comme VIAF ou Wikidata.

Les données venant alimenter l'instance Wikibase incluent à la fois un chargement initial de jeux de données nominatifs du CegeSoma et d'autres données – sur des personnes liées aux collections du CegeSoma ou des Archives de l'État – ajoutées [197] en continu par les personnnes de l'institution participant à l'encodage des données, mais aussi potentiellement par des chercheurs invités à partager les données de leurs recherches. Enfin, les entités Personne créées dans l'instance Wikibase sont potentiellement connectées par des liens d'équivalence à des entités externes issues d'autres bases de connaissance comme Wikidata.

197. Le chapitre 5 abordera plus en détail la façon dont ces données peuvent être concrètement ajoutées.

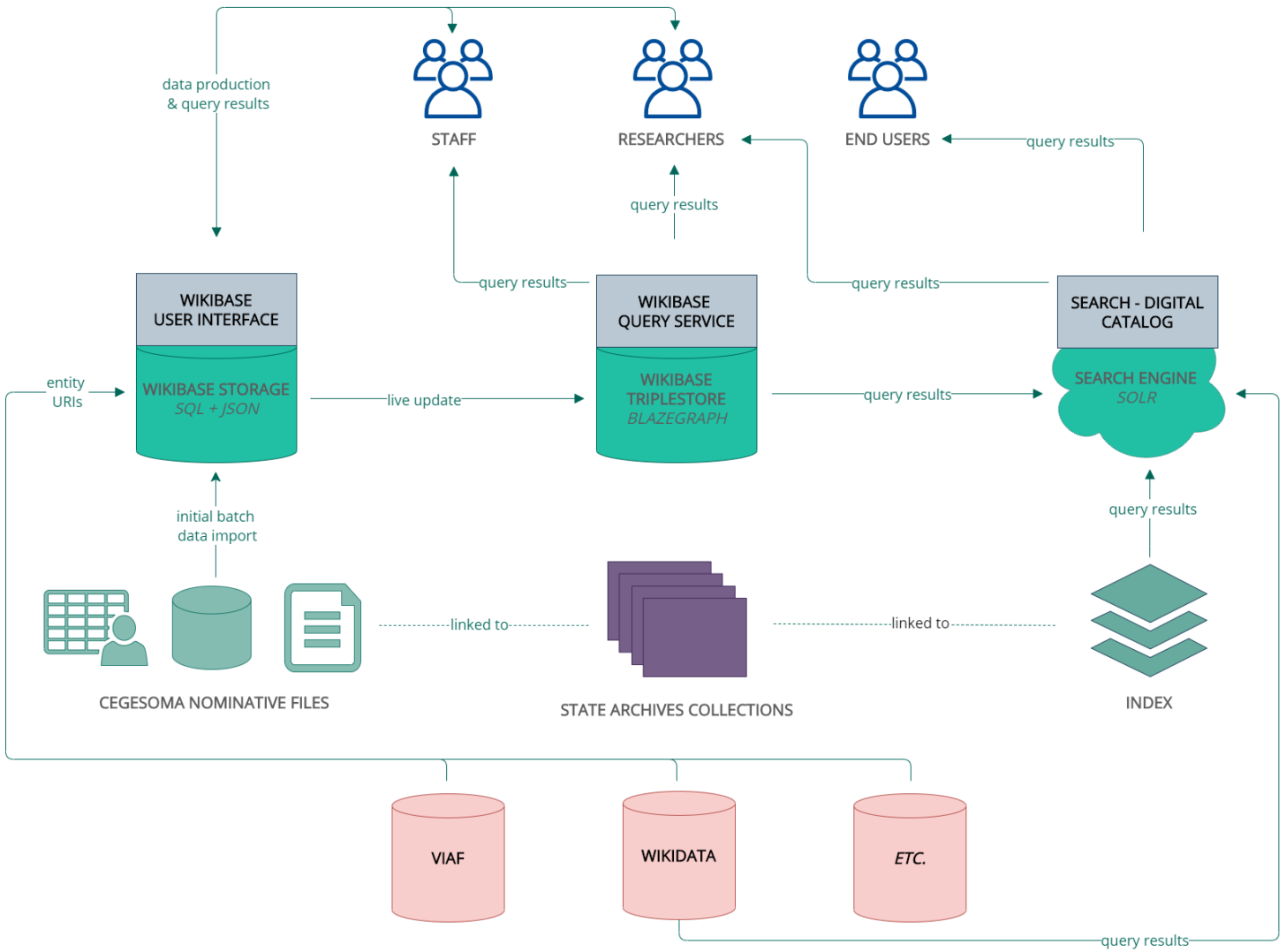


FIGURE 2.11 – Vue d'ensemble des interactions prenant place dans le cadre de l'implémentation d'une instance Wikibase.

3 | Contexte

Introduction

Ce chapitre introductif, dédié au contexte, est composé de deux sections. La première section présente l'institution dont est issue cette étude de cas, le Centre d'Études et de Documentation Guerre et Sociétés contemporaines (CegeSoma), en explicitant notamment les importants changements auxquels l'institution a été confrontée au cours des dernières années. En effet, le CegeSoma a été intégré aux Archives de l'État en Belgique le 1er janvier 2016. Or, cette fusion entraîne dans son sillon de multiples questions concernant le système de gestion et de description des collections, en particulier le langage d'indexation-matière utilisé depuis plus de 20 ans par le Centre pour décrire ses archives, manuscrits, livres et brochures, périodiques, photographies, tracts, affiches, coupures de presse, interviews et matériaux audiovisuels. Les Archives de l'État ne proposant – pour l'instant – pas de système de gestion et d'affichage de termes concernant les sujets traités par les documents, la question se pose par exemple de savoir ce qui va advenir de ces 120 000 descripteurs contenant un nombre important de noms propres. Ce questionnement entre par ailleurs en résonance avec une réflexion plus large menée par le CegeSoma sur la gestion des métadonnées et les opportunités offertes par les *Linked Data*, alors même que le personnel continue d'encoder dans des fichiers isolés des informations sur des milliers d'individus associés à l'un ou l'autre fonds d'archives. Nous allons ainsi voir au cours des pages à venir comment ce qui apparaissait à première vue comme une difficulté peut être envisagé comme une opportunité.

La seconde section s'intéresse aux besoins de l'institution et de ses utilisateurs en ce qui concerne l'accès numérique aux collections, avec une emphase mise sur l'indexation et les données d'autorité. En effet, cette thèse mobilise le principe de *fitness for use*, un principe inspiré de la norme ISO 9001 pour le management de la qualité. La version de 2005 de cette norme définit la qualité comme :

> The totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs (ISO, 2005).

Cette définition permet ainsi de pallier l'absence de référence absolue auquel doit faire face le secteur du patrimoine culturel. En effet, comme l'ont souligné Boydens et van Hooland, le caractère empirique des métadonnées et l'absence de référence *absolue*, rendent difficile l'évaluation de la qualité des métadonnées à un niveau intellectuel (Boydens et van Hooland, 2011). Dans le cas des données d'autorité du CegeSoma, ce principe de *fitness for use* nous invite à nous intéresser au contexte et à relever les besoins implicites et explicites des parties prenantes qui sont confrontées à ces données [1]. Inspirée du rapport du groupe de travail de l'IFLA (Fédération internationale des associations et institutions de bibliothèques) sur les Fonctionnalités requises et la numérotation des notices d'autorité (FRANAR), qui identifie comme utilisateurs des données d'autorité [2] à la fois « les créateurs des données d'autorité qui créent et maintiennent les données » et « les utilisateurs qui se servent des informations contenues dans les autorités [...] » (FRANAR, 2010), cette section propose donc de s'intéresser tant à l'institution et à son personnel qu'aux utilisateurs du CegeSoma. Une fois ces besoins identifiés, une solution permettant au Centre d'améliorer sa gestion des données d'autorité est esquissée.

3.1 Le CegeSoma

3.1.1 Présentation

Le CegeSoma est le centre d'expertise belge de l'histoire des conflits du XX^{e} siècle. Sa naissance remonte au 13 décembre 1967, sous la dénomination de Centre de Recherches et d'Études historiques de la Seconde Guerre

1. Cependant, comme le souligne l'archiviste du Centre qui s'inscrit dans la ligne de pensée de l'archiviste Gerald Ham par rapport à la politique d'acquisition, il faut toutefois être vigilant et éviter de forger une politique visant uniquement l'utilisateur final actuel : « De maatschappelijke opdracht van onze instelling overstijgt evoluerende methodologieën en theoretische kaders, onderzoekshypes of actuele brandpunten. Als archiefcentrum heeft het CegeSoma de opdracht om een collectie samen te stellen die de Belgische conflictgeschiedenis in al haar facetten belicht. Het CegeSoma moet dus durven archieven en andere stukken te verwerven zelfs als ze niet onmiddellijk aantrekkelijk blijken of bruikbaar zijn voor de onderzoekers van vandaag. ». (Desmet, 2019, p. 9)

([Traduction libre] La mission sociale de notre institution transcende les méthodologies et les cadres théoriques en évolution, les recherches en vogue ou les centres d'intérêt actuels. En tant que centre d'archives, le CegeSoma a pour mission de constituer une collection mettant en lumière l'histoire des conflits belges sous toutes ses facettes. Le CegeSoma doit donc oser acquérir des archives et d'autres types de ressources même s'ils ne sont pas immédiatement attrayants ou utiles pour les chercheurs d'aujourd'hui).

2. Ici, dans le contexte des bibliothèques.

mondiale (CREHSGM). Il est appelé à « prendre toutes les mesures nécessaires en vue de recenser, sauvegarder et dépouiller les documents ou archives se rapportant à la Seconde Guerre mondiale en Belgique, à ses antécédents et préliminaires ainsi qu'à ses conséquences »[3]. Rattaché aux Archives générales du Royaume et placé sous la tutelle du Département de l'Éducation nationale, il doit attendre l'été 1969 pour disposer de ses propres locaux, rue Joseph II, à Bruxelles (Colignon, 2019). Sous la direction de Jean Vanwelkenhuyzen, une équipe composée d'une secrétaire, d'un commis et de quatre chercheurs s'attèle à l'étude de la Seconde Guerre mondiale, mais également à la constitution d'une bibliothèque et à la collecte d'archives et de photographies liées à la période 1939-1945 (Gotovitch, 2005). Il faut attendre 1972 pour qu'une première salle de lecture destinée à l'accueil du public ouvre ses portes, et, dès les années 1980, le Centre met sur pied des séminaires mensuels, suivis plus tard par l'organisation de colloques internationaux. Peu à peu l'équipe s'agrandit et voit la nomination d'un nouveau directeur en 1989, José Gotovitch.

Dans les années 1990, l'institution se lance dans divers chantiers informatiques : la numérisation de ses collections, la création d'un thésaurus informatisé et l'engagement d'un informaticien à temps plein (Colignon, 2019). En 1997, alors que l'institution est rebaptisée et devient le Centre d'Études et de Documentation Guerre et Sociétés contemporaines (CEGES), elle se dote d'un système informatisé de gestion et de description des collections, désigné sous le nom de Pallas[4]. Puis, dès l'an 2000, le CEGES publie son propre site web[5].

En 2005, José Gotovitch cède sa place à Rudi Van Doorslaer. En 2011, les secteurs Documentation et Activités académiques sont rejoints par un secteur Histoire publique, qui sera suivi par le secteur Digitalisation en 2016. Cette année 2016 voit également l'intégration du Centre, désormais connu sous le nom de CegeSoma, aux Archives de l'État en Belgique, dont il constitue la quatrième Direction opérationnelle (D04) (CegeSoma, 2020a). Cette intégration se traduit par l'adoption progressive de nouveaux outils de travail et la migration progressive des descriptions des collections du CegeSoma vers le système de gestion des Archives de l'État. 2016 marque également le départ à la retraite de Rudi Van Doorslaer, remplacé par Dirk Martin, puis par Fabrice Maerten comme directeur *ad interim* jusqu'au 1er mai 2017, date

3. Extrait du Moniteur belge (10 février 1968, numéro 29, pp. 1259-1261) cité par (CegeSoma, 2020a).

4. Développé en interne par l'informaticien Patrick Temmerman, il s'appuie sur le système de gestion de base de données Oracle.

5. Une archive du premier site de l'institution est disponible en ligne : https://web.archive.org/web/20000529183959/http://www.cegesoma.be/

à laquelle Nico Wouters entre en service comme chef fonctionnel du CegeSoma.

La quatrième Direction opérationnelle des Archives de l'État compte aujourd'hui une vingtaine de collaborateurs, parmi lesquels huit collaborateurs scientifiques permanents. Elle a par ailleurs la chance de pouvoir compter sur le soutien d'une petite dizaine de bénévoles, d'étudiants jobistes et de stagiaires [6]. Le fonctionnement de l'institution est organisé autour de sept grands axes : digitalisation ; bibliothèque et informations au public ; valorisation des collections ; communication et événements ; recherche ; histoire publique ; gestion des archives (CegeSoma, 2020b).

Les missions du Centre sont décrites ainsi sur le nouveau site web du CegeSoma [7] :

> Le Centre d'Etude Guerre et Société est le centre belge d'expertise sur l'histoire des conflits du XX^{e} siècle. L'institution constitue depuis le 1er janvier 2016 la 4e Direction opérationnelle (DO4) des Archives de l'Etat [...]. Ses centres d'intérêt sont principalement liés à la Première et à la Seconde Guerre mondiale, y compris à leur long impact sociétal et à leur héritage. Ses tâches principales sont la recherche scientifique, l'histoire publique, la documentation et la numérisation. (CegeSoma, 2020b)

La comparaison de ces quelques lignes avec le *mission statement* (CegeSoma, 2019a) diffusé sur l'ancien site web du CegeSoma témoigne du resserrement des thématiques couvertes par le Centre. Ce document, datant de juin 2007, expliquait alors que :

> [Le Ceges] s'assigne pour objectif d'être un centre belge de référence pour l'histoire des grands conflits et moments de rupture sur les plans politique, social et culturel au XX^{e} siècle. (CegeSoma, 2007)

Si les deux guerres mondiales y occupaient déjà une place centrale, les autres conflits et idéologies totalitaires caractérisant le XX^{e} siècle complétaient alors ses centres d'intérêt. La nouvelle mouture, élaborée en 2016 lors de l'intégration du Centre aux Archives de l'État, témoigne finalement d'un désir de l'institution de renouer avec ses fondamentaux et de réaffirmer sa spécificité, en se concentrant sur ses missions essentielles (CegeSoma, 2018a). Cette spécificité tient également au fait que, contrairement aux autres directions opérationnelles des Archives de l'État, le Centre ne développe pas ses activités sur une logique géographique (Desmet, 2019, p. 3) et que sa

6. Un soutien qui est par ailleurs loin d'être anecdotique : il correspond à plus de 3 700 heures prestées pour l'année 2018 (CegeSoma, 2019b, p. 21).

7. Lancé en décembre 2019.

politique d'acquisition est beaucoup plus proche de celle du secteur des archives privées que de celles des autres dépôts des Archives de l'État, tenus de surveiller et d'organiser les transferts d'archives de droit public (Desmet, 2019, p.8).

Le resserrement du Centre vers ses thématiques de base est également une conséquence de la diminution de ses moyens au cours des dernières années. En effet, il semble important de terminer cette présentation en mentionnant les variations importantes des moyens dont le Centre a pu disposer. Il nous est apparu que l'évolution de la bibliothèque de l'institution, récemment consignée sur papier par son responsable actuel – Alain Colignon – constitue une précieuse source d'information. En effet, si cette bibliothèque constitue un véritable reflet des « extensions ponctuelles sur le plan chronologique ou conceptel » et de la « diversification des missions » qu'a connus le Centre au gré de ses différents directeurs, elle témoigne également de l'instabilité des moyens budgétaires dont bénéficie le Centre (Colignon, 2019). Le bibliothécaire du Centre nous apprend ainsi que durant un mandat débuté en 2005, l'ancien directeur Rudi Van Doorslaer était guidé par l'ambition de transformer la bibliothèque du Centre en *Bibliothèque d'Histoire du* XX*e siècle*, voire même en *Bibliothèque universelle du Temps présent*. Il n'hésita dès lors pas à modifier le budget qui lui était alloué en conséquence, permettant ainsi à la bibliothèque d'accroître son stock, ce dernier passant de 50 000 à 60 000 ouvrages [8] en l'espace de trois ans.

Concrètement, si le budget annuel de la bibliothèque gravitait autour de 20 000 euros par an (Colignon, 2019, p. 156) dans les années 2002-2005, il grimpa jusqu'à 50 000 euros entre 2008 et 2012, avant que les choses ne se mettent à décliner graduellement : « les sommes allouées à la bibliothèque allèrent se réduisant comme peau de chagrin au fur et à mesure des tours de vis successifs pratiqués, à partir de 2014, par l'équipe ministérielle en place » (Colignon, 2019, p. 156). Ainsi, bien que la bibliothèque put encore compter sur un budget de 13 000 euros en 2016, la rigueur budgétaire s'est poursuivie, amenant le responsable de la bibliothèque du CegeSoma a tiré la sonnette d'alarme : le budget initialement prévu [9] était alors de 5 000 euros (Colignon, 2018), soit 10% seulement du budget alloué six ans plus tôt.

Or cet affaissement budgétaire touche l'institution dans son ensemble et pas uniquement le secteur de la bibliothèque : ainsi, entre 2016 – qui marque le début notre collaboration avec le CegeSoma, dans le cadre du projet Adochs – et 2019, nous avons été témoin du départ de plusieurs collaborateurs dont la plupart n'ont pu être remplacés, alors même que le Centre

8. Pour la série générale.

9. Au final, un apport inattendu de 7 000 euros, issu des budgets résiduels d'anciens projets de recherche, vint compléter la somme ; il s'agit cependant d'un acte ponctuel qui ne peut être considéré comme une source régulière et pérenne de financements.

comptait une cinquantaine de collaborateurs dix ans plus tôt [10]. En effet, le CegeSoma, à l'instar des établissements scientifiques fédéraux belges (Patrick *et al.*, 2014), s'avère soumis à des coupes budgétaires le condamnant à devoir rechercher des moyens alternatifs [11] pour tenter de poursuivre ses activités [12]. Comme nous le verrons, si ce contexte marqué par des ressources limitées entraîne des questionnements quant à la pérennité et la maintenance des outils mis en place, il donne également une résonnance particulière aux dimensions de mutualisation et de fédération que permet le Web de données.

3.1.2 Collections

Depuis mai 2018, le CegeSoma peut compter sur la présence d'un archiviste statutaire. L'arrivée de ce nouveau profil – venu enrichir une équipe scientifique constituée exclusivement d'historiens et d'historiennes jusque-là – correspond à un moment charnière dans les efforts du Centre pour professionnaliser sa gestion des collections : comme l'explique le rapport annuel de 2018, cette présence constitue un préalable à la mise en œuvre d'une « réforme cruciale » allant de la gestion des dépôts à l'ouverture à la recherche numérique des collections (CegeSoma, 2019b, p. 4). Cette réforme, qui se traduit notamment par l'énonciation d'une politique claire d'acquisition des collections, permettant de rendre plus cohérente et explicite une politique caractérisée jusque-là par son caractère informel (Desmet, 2019, p. 7), fut l'occasion de tracer explicitement les contours des collections du Centre :

> Het CegeSoma verzamelt private archieven, iconografische en audiovisuele bronnen, alsook bibliothecaire collecties die betrekking hebben op 20ste-eeuwse conflictgeschiedenis, met nadruk op de Eerste en Tweede Wereldoorlog en de Koude Oorlog (incl. voor- en nageschiedenis), voor wat betreft het grondgebied van de Belgische staat met inbegrip van haar voormalige koloniale bezittingen in Midden-Afrika [13]. (Desmet, 2019, p. 7)

10. Entretien avec Alain Colignon, 20.01.2020.

11. Comme la recherche de sources de financements externes par le biais de projets de recherche (CegeSoma, 2018a), ou le recours à des stagiaires et bénévoles pour rédiger des inventaires (CegeSoma, 2020c).

12. Il faut toutefois nuancer cela : ainsi, il s'avère que l'un des anciens directeurs du Centre, José Gotovitch, déplorait lui-même ces « temps de restriction budgétaire » en 2004 (Gotovitch, 2005).

13. [Traduction libre] Le CegeSoma rassemble des archives privées, des sources iconographiques et audiovisuelles, ainsi que des collections de bibliothèques relatives à l'histoire des conflits du XX^{e} siècle, en mettant l'accent sur la Première et la Seconde Guerre mondiale, ainsi que la Guerre froide (y compris l'histoire d'avant et d'après-guerre) en ce qui concerne les territoires de l'État belge, y compris ses anciennes possessions coloniales en Afrique centrale).

Par ailleurs, il faut noter que ces collections sont destinées à n'accueillir que des documents de droit privé, tels que des archives, des écrits autobiographiques ou des photographies issues d'entités privées telles que des personnes et des familles, des associations ou des sociétés (Desmet, 2019, p. 12,14). Ce qui confère à l'institution une autre de ses particularités, à savoir le caractère social que revêt son activité de conservation de documents d'archives :

> Door [het] archief van particulieren en private organisaties te bewaren en te valoriseren neemt het Centrum deel aan het faciliteren van historisch onderzoek over, en de constructie van de collectieve herinnering(en) aan de twintigste-eeuwse conflicten. Op dit vlak kan het CegeSoma binnen het Rijksarchief een voortrekkersrol spelen [14]. (Desmet, 2019, p. 10)

Pour faciliter l'accès à cette recherche historique et contribuer à la mise en accès de son patrimoine archivistique, le Centre s'est lancé dès 1997 dans la numérisation de ses collections, en se concentrant d'abord sur ses archives photographiques, qui font l'objet d'une forte demande (Gillet, 2019, p. 3). En 2019, l'ensemble des collections photographiques [15] est accessible sous forme numérique, de même que la presse de guerre, diverses transcription d'interviews et d'émissions radiophoniques, certains fonds d'archives tels que les archives von Falkenhausen-Canaris, une partie des affiches et tracts, des périodiques concernant la collaboration, ainsi que des archives sonores telles que les émissions Radio Bruxelles-Zender Brussel (Gillet, 2019, p. 8-9).

Pour gérer ses collections, le CegeSoma a longtemps fait usage d'un logiciel développé en interne : Pallas. Ce dernier, créé sur mesure pour répondre aux besoins du Centre et lancé en 1997, a le mérite de pouvoir prendre en charge – en suivant les normes internationales de description telles que ISAD(G), MARC et SEPIADES – des collections variées : du tract au manuscrit, en passant par les fonds d'archives, les affiches et les collections photographiques, sonores et audiovisuelles. Il a en outre la particularité de cumuler différentes fonctions : il fait à la fois office d'outil de gestion, de description et d'interface de consultation des collections (Gillet, 2019). Mais entre le fait que ce système de gestion documentaire n'était plus à même de répondre aux besoins des utilisateurs (Gillet, 2019) et la récente intégration de l'institution aux Archives de l'État, le Centre a dû se résoudre à abandonner Pallas. À l'heure actuelle – printemps 2020 –, la migration des données

14. [Traduction libre] En préservant et en valorisant les archives de particuliers et d'organisations privées, le Centre participe à la facilitation de la recherche historique et à la construction de la mémoire collective des conflits du XX^{e} siècle. En cela, le CegeSoma peut jouer un rôle de pionnier au sein des Archives de l'Etat.

15. Soit environ 350 000 photos.

vers le système de gestion des Archives de l'État est entamée et les collections du Centre sont appelées à être consultées à l'avenir à l'aide du moteur de recherche des Archives de l'État : Search [16], qui est actuellement en cours de refonte.

3.1.3 Gestion des métadonnées

Les changements entraînés par l'intégration de l'institution aux Archives de l'État et la migration de ses données signifient que le personnel doit progressivement s'adapter à de nouvelles procédures de travail, notamment en ce qui concerne la gestion des métadonnées. Au vu de notre objet d'étude, nous prenons le temps ici d'aborder cette question plus en détail, et plus précisément sous l'angle de la dette technique, un concept présenté au cours des premières pages de cette thèse. Nous nous inspirons ici de la typologie proposée par Clair (2016) pour évaluer les différents types de dettes pouvant être distingués dans le cadre de la gestion des métadonnées d'une bibliothèque. Elle comprend cinq dimensions :

1. Code
2. Design et architecture
3. Environnement
4. Documentation
5. Exigences [17]

Le tableau 3.1 propose une vue récapitulative de ce que recouvrent ces cinq dimensions dans trois contextes différents : premièrement, dans le contexte du développement logiciel [18], deuxièmement, dans le contexte des bibliothèques [19] et enfin, dans le contexte plus spécifique du CegeSoma [20].

Cette analyse des différents types de dette pouvant toucher les métadonnées nous permet d'esquisser un portrait des différents défis auxquels le CegeSoma est actuellement confronté. Le premier type de dette, lié au code, permet par exemple de mettre en lumière les difficultés occasionnées par une inadéquation entre le mode d'encodage des métadonnées au sein du logiciel Pallas et les standards en vigueur. Des problèmes d'interopérabilité [21]

16. http://search.arch.be/.
17. Dans la littérature, c'est le terme anglais *requirements* qui est utilisé.
18. Définitions provenant des typologies de Tom *et al.* (2013) et de Li *et al.* (2015), telles que reprises par Clair (2016).
19. Définitions provenant des travaux de Clair (2016).
20. Sur la base d'échanges avec la responsable de l'accès numérique aux Collections.
21. En raison de l'utilisation d'une version EAD obsolète au sein de Pallas et d'un système de notation de dates ne répondant pas aux standards requis par EHRI.

se sont ainsi présentés lorsque le CegeSoma a souhaité publier ses inventaires EAD sur la plateforme internationale EHRI (European Holocaust Research Infrastructure) [22], mais également lorsqu'il a été question de l'exportation de fonds d'archives depuis le logiciel Pallas, certains fonds ayant été erronnément décrits au niveau du dossier. Par ailleurs, il faut également relever divers problèmes de qualité des métadonnées incluant des défaillances en matière de sémantisation, c'est-à-dire tout un arriéré de données non structurées, encodées en langage naturel et dès lors sujettes à l'ambiguïté.

Le deuxième type de dette concerne les dimensions de design et architecture. Tout d'abord, il faut relever que si le logiciel Pallas fut considéré comme un précurseur à ses débuts [23], la version utilisée au CegeSoma est aujourd'hui dépassée et ne prend par exemple pas en charge ISAAR(CPF), la norme internationale sur les notices d'autorité archivistiques relatives aux collectivités, personnes et familles. Cette dette est renforcée par le fait que le CegeSoma est actuellement occupé, comme souligné au cours des pages précédentes, à organiser la migration de ses données vers un nouveau système de description des collections (SAM), ainsi que vers un nouveau système de découverte des collections – par ailleurs en cours de refonte. Ces changements au niveau du système de gestion des métadonnées ont bien entendu un impact direct sur les métadonnées. D'une part, parce que cela occasionne des opérations de mise en correspondance pouvant passer par l'utilisation d'un système transitoire accroissant le risque d'erreurs et de pertes de données [24], d'autre part, parce que le nouveau système ne prend pas en charge l'intégralité des métadonnées utilisées jusque-là : c'est le cas par exemple des descripteurs du système d'indexation-matière utilisé par le CegeSoma depuis plus de 20 ans.

La troisième dimension que peut prendre cette dette est liée à l'environnement dans lequel s'inscrit la gestion des métadonnées. Dans le cadre du CegeSoma, le fait que l'institution ne dispose désormais plus directement d'informaticien en interne [25], semble renforcer les difficultés de compréhension pouvant survenir dans le cadre de la communication entre le personnel

22. https://www.ehri-project.eu/.

23. Il intègre ISAD(G), un standard à l'époque « encore méconnu » (Temmerman, 2007, p. 24) et permet au CegeSoma d'être « l'une des plus grandes bases de données documentaires d'Europe gratuitement consultable » et la seule à intégrer à la fois des archives et des images (Temmerman, 2007, p. 25). Par ailleurs, dès 2004 le logiciel est adopté par d'autres institutions (centres de documentation, bibliothèques spécialisées ou dépôts d'archives) et la Communauté française décide d'en faire l'acquisition afin de pouvoir en équiper les centres d'archives qu'elle subsidie (Temmerman, 2007, p.24).

24. Les problèmes d'export rencontrés avec le logiciel Pallas ont ainsi conduit le personnel à opter pour l'utilisation d'un fichier Excel rempli à l'aide d'opérations de copié-collé, faute de mieux.

25. Elle dépend désormais des services d'appui communs à l'ensemble des Archives de l'État.

Dimensions	Développement logiciel	Bibliothèques	CegeSoma
1. Code	Code mal rédigé, illogique, redondant ou en contradiction avec des règles existantes	Métadonnées erronées, incomplètes ou en contradiction avec les règles de catalogage (cette dette peut être tant intentionnelle qu'involontaire)	Inadéquation entre certains modes d'encodage des données et les standards en vigueur, problèmes de qualité
2. Design et architecture	Code rédigé pour satisfaire les besoins d'un certain type d'interface, au détriment d'une vision à long terme	Décisions déterminées par le système de découverte des collections et, ou conversion vers un nouveau système et, ou modification des normes de description	Logiciel obsolète, migration des métadonnées vers de nouveaux systèmes de description et de découverte des collections
3. Environnement	Manque d'efficacité au niveau de l'environnement et des processus menant au développement du code	Défaillances dans la communication entre les services techniques et les autres services de l'institution	Défaillance dans la communication entre services, ressources limitées, absence de politique claire de gestion des métadonnées
4. Documentation	Documentation insuffisante, incomplète ou dépassée du code	Documentation insuffisante des normes et des règles locales complétant ces normes	Gestion des connaissances lacunaire, documentation insuffisante du système de gestion des collections, Pallas
5. Exigences	Écart entre les exigences initiales et la mise en œuvre effective du système	Écart entre les divers cas d'usage qu'un système de gestion des métadonnées est sensé prendre en charge et sa mise en œuvre effective	Écart entre les besoins des utilisateurs et ce que permet le catalogue en ligne Pallas

TABLE 3.1 – Vue récapitulative des différentes dimensions de la dette technique dans le contexte du développement logiciel et des bibliothèques (Clair, 2016), ainsi que dans le contexte du CegeSoma.

chargé de la gestion des collections et les services ICT. Ces difficultés sont elles-mêmes amplifiées par des divergences de pratiques descriptives entre le CegeSoma et les Archives de l'État. De plus, il faut noter que l'environnement est également marqué par un important taux de renouvellement du personnel, ainsi que par des ressources limitées se soldant par des processus de travail non optimaux [26]. Enfin, l'absence de politique claire de contrôle des métadonnées et le manque d'harmonisation à ce sujet au sein de l'institution tendent à exarcerber les problèmes de qualité.

Quatrièmement, la dette est alourdie par l'absence de documentation suffisamment à jour et complète, ainsi que par des lacunes au niveau de la gestion des connaissances. En effet, si le Centre – désormais dans le giron des Archives de l'État – dispose aujourd'hui d'un véritable système de *Records Management* qui a permis de systématiser le partage de fichiers de documentation, ce n'était pas encore le cas lors du départ de l'institution de l'informaticien [27] à l'origine du logiciel de description des collections, Pallas. Si ce manque de documentation n'a pas empêché l'institution de continuer à utiliser ce logiciel bien après le départ de son créateur, en revanche il s'est avéré critique dans le contexte de la migration des données vers le système des Archives de l'État. L'équipe chargée de la migration des données de Pallas a ainsi dû commencer par entreprendre d'importantes opérations de *reverse engineering* de manière à pouvoir comprendre l'utilité et les particularités de l'ensemble des tables utilisées.

Enfin, la cinquième dimension de la dette naît de l'écart entre les besoins et les attentes des différentes types d'utilisateurs du système documentaire et les fonctionnalités réellement prises en charge par ce dernier. Comme nous le verrons en détail au cours des prochaines pages, le CegeSoma s'est donné comme objectif d'étudier ces besoins au cours des dernières années, aboutissant à des constats venant mettre en lumière les inévitables limites d'un logiciel conçu voilà plus de 20 ans.

3.2 Analyse des besoins

3.2.1 Besoins de l'institution

Bien que le personnel de l'institution puisse être considéré comme faisant partie des utilisateurs au sens large, nous tenions à d'abord prendre le temps de passer en revue les besoins de l'institution – exprimés de façon plus ou

26. C'est le cas par exemple lorsqu'une personne dotée d'une forte expertise est contrainte à réaliser des opérations chronophages d'encodage faute de personnel disponible, au détriment d'opérations plus stratégiques.

27. Qui nous a par ailleurs été décrit comme un informaticien autodidacte n'ayant pas systématiquement suivi les standards informatiques.

moins explicite –, avant de passer au reste des utilisateurs dans une prochaine sous-section. Plusieurs ressources ont été consultées dans le but de pouvoir identifier ces besoins : les ressources consultées comprennent à la fois des documents de travail circulant en interne et des publications plus officielles telles que des rapports d'activité. À défaut de disposer d'un rapport approfondi portant sur les données d'autorité en particulier, nous nous sommes intéressée aux priorités formulées par le personnel pour les années à venir et aux besoins que cela donne à voir en filigrane. Les prochains paragraphes mettent ainsi en exergue des informations relatives à la stratégie globale du CegeSoma, à l'accès aux collections, aux outils de recherche, au déploiement du numérique, ainsi qu'aux projets axés *digital humanities*.

En juin 2017, un rapport d'audit *(peer review)* réalisé aux Archives de l'État était publié. L'une des recommandations concernait spécifiquement le rôle et la place du CegeSoma :

> Keep CegeSoma visible within the ARA structure and build on the Centre's original mission and acquired expertise. This means a focus on the core expertise of the history and impact of war and mass violence on Belgian contemporary history. (Zuijdam *et al.*, 2017, p. 7)

C'est poussé par ce désir de renforcement de ses missions essentielles et par une consolidation de son expertise que le Centre a initié une réorganisation interne et la publication d'un plan pluriannuel pour la période 2018-2021. Ce dernier, basé sur ses missions-clés [28], dresse une liste de cinq objectifs stratégiques (CegeSoma, 2018a) :

1. Le CegeSoma veut améliorer la communication d'informations et le service de qualité sur le plan qualitatif (numérique et traditionnel) portant sur ses missions essentielles.
2. Le CegeSoma se concentre sur de nouvelles formes de recherche fondamentale relative à ses missions essentielles. Il renforce sa position académique aux niveaux national et international.
3. Le CegeSoma se concentre sur l'histoire publique et le débat sociétal.
4. Le CegeSoma se concentre sur une valorisation qualitative (numérique et traditionnelle) des collections des AGR relatives à ses missions essentielles.
5. Le CegeSoma se concentre sur la gestion, le développement et l'ouverture à la recherche des collections.

28. Ces missions, confirmées par l'Arrêté royal du 23 mai 2016 ayant fait du CegeSoma la quatrième Direction opérationnelle des Archives de l'État (CegeSoma, 2018b), sont détaillées dans les paragraphes dédiés à la présentation de l'institution.

Comme le souligne Florence Gillet – responsable de l'accès numérique aux collections et des projets en Humanités numériques au CegeSoma –, le numérique est présent dans chacun de ces objectifs stratégiques (Gillet, 2019). C'est toutefois le quatrième et le cinquième objectifs qui retiennent le plus notre attention et semblent se recouper sur certains aspects. La concrétisation du quatrième objectif est résumée ainsi :

> En tant que centre gestionnaire de collections avec des missions essentielles en matière de recherche et d'histoire publique, le CegeSoma doit s'investir dans la mise en œuvre d'instruments d'accès à la recherche [...] tant classiques que sous des formes nouvelles [...]. L'objectif essentiel est de prendre en compte les besoins de la recherche, de la digitalisation et de l'histoire publique des utilisateurs externes mais aussi de l'institution. (CegeSoma, 2018a, p. 4)

Il est également précisé que le Centre est appelé à proposer de nouvelles formes de coopération au sein des AGR (Archives Générales du Royaume) et que, fort de son expérience issue de projets de valorisation des collections numériques, comme le projet UGESCO [29] :

> Le CegeSoma souhaite acquérir une position spécifique au sein des AGR dans le domaine des humanités numériques en se focalisant sur des projets de recherche et des projets d'histoire publique en vue de valoriser les collections des AGR. (CegeSoma, 2018a, p. 4)

Quant au cinquième objectif, qui implique une politique ciblée en matière d'inventorisation des collections et d'acquisition d'archives privées, il se traduit également par la mise en place d'instruments numériques pour la diffusion interne des informations, ainsi que par :

> L'optimalisation de l'accès numérique aux collections par le grand public [qui] nécessite aussi une amélioration des plateformes existantes telle War Press [30] et PALLAS, notamment via une meilleure gestion des métadonnées et une convivialité accrue des interfaces existantes. (CegeSoma, 2018a, p. 5)

En parallèle de la publication de ce plan pluriannuel, le nouveau directeur opérationnel a également invité ses collègues membres du personnel scientifique permanent à rédiger des notes de vision. C'est le cas notamment

29. Le projet UGESCO (Upscaling the Geo-temporal Enrichment, exploration and exploitation of Scientific Collections) s'est déroulé de 2017 à 2019 et a permis de travailler sur les métadonnées spatiotemporelles des collections photographiques du CegeSoma, voir https://www.cegesoma.be/fr/project/ugesco.

30. Ce site, mis en ligne par le CegeSoma, propose un accès à la presse de guerre numérisée, voir : https://warpress.cegesoma.be/.

de la responsable du numérique au CegeSoma, qui souligne que le Centre se situe à un tournant numérique. Elle évoque alors la nécessité de remplacer son système de gestion documentaire ; l'intégration de l'institution dans les Archives de l'Etat ; le développement du numérique ces dernières années dans le secteur patrimonial et une expertise incontestable en Histoire publique numérique et en humanités numériques. Relevant les points forts du Centre (l'intégration du Centre comme quatrième Direction opérationnelle des Archives de l'État, la meilleure connaissance des besoins des utilisateurs issue du projet MADDLAIN et les contacts développés avec les universités), elle note également plusieurs obstacles : le manque de moyens conduisant le Centre à dépendre de moyens ponctuels dépendant de financements extérieurs, « compliqu[ant] considérablement le développement d'une stratégie à long terme ainsi que le maintien des compétences humaines nécessaires en interne » ; les divergences de pratiques avec les Archives de l'État complexifiant l'intégration des données du CegeSoma ; la centralisation des ressources informatiques du CegeSoma vers le service ICT des Archives de l'État (Gillet, 2019, pp. 5-6).

Tenant compte de cela, mais également des moyens disponibles et des attentes exprimées par les différents publics concernés, elle propose ainsi trois axes de travail :

1. L'accès numérique aux collections
2. La création de données numériques de qualité
3. La mise à disposition de données ouvertes pour la recherche

Si l'accès numérique aux collections (premier axe de travail) concerne principalement la migration des données du Centre vers les outils de gestion des Archives de l'État, ce sont les deuxièmes et troisièmes axes de travail qui retiennent particulièrement notre attention dans le cadre de cette thèse sur les données d'autorité. Le deuxième axe de travail est introduit ainsi :

> Le développement du numérique dans une institution patrimoniale passe inévitablement par la création de données numériques de qualité. En d'autres mots, il s'agit de mettre à disposition des chercheurs des données correctes (dont le contenu a été vérifié), normalisées (qui respectent les standards utilisés au niveau international) et interopérables (de manière à pouvoir les échanger facilement). Mettre en place une réelle stratégie de gestion des données numériques, c'est non seulement assurer leur pérénité mais également faciliter l'accès des lecteurs aux collections, rendre possible leur interconnexion avec le patrimoine conservé dans d'autres institutions sœurs et leur utilisation par

> de multiples acteurs. Ce constat vaut tant pour les métadonnées que pour les fichiers numériques. (Gillet, 2019, p. 15)

Nous retenons plus particulièrement ici la nécessité de pouvoir bénéficier de données correctes – devant dès lors être vérifiables à l'aide d'outils appropriés –, normalisées et interopérables.

Quant au troisième axe de travail, concernant l'ouverture des données numériques à la recherche et le développement de projets *Digital Humanities*, Gillet explique que le CegeSoma dispose de deux possibilités. Premièrement, la mise en ligne de données brutes à disposition des chercheurs tels qu'un fichier Excel contenant les légendes des collections photographiques ou les fichiers PDF des inventaires. La seconde possibilité s'offrant au CegeSoma est de participer au développement de projets de recherches *Digital Humanities* basés sur les collections du Centre. Or, ce type de projets soulève la question de la qualité des données numériques et de leur mise à disposition. Parmi les actions concrètes envisagées pour avancer dans cette direction, Gillet cite notamment « le développpement d'une stratégie Linked Open Data pour améliorer l'accès aux collections du CegeSoma et des autres DO des Archives de l'État via le projet ADOCHS » (Gillet, 2019, p. 19).

Parmi les notes de vision rédigées par les autres responsables de secteur du CegeSoma, nous relevons une certaine superposition de contenu. En effet, bien que l'organigramme de l'institution prévoie un fonctionnement par secteur, beaucoup d'éléments se recoupent, comme le relève par exemple la note d'intention du responsable de la valorisation des collections, « [cette] matière transversale qui concerne pratiquement tous les scientifiques du CegeSoma » (Maerten, 2019, p. 1). Ainsi, les quatre grands axes de travail cités pour ce secteur [31] apparaissent également dans les autres notes : l'acquisition et l'ouverture à la recherche ; l'amélioration de l'accès via des outils numériques ; la communication au public ; les *produits* de valorisation. Parmi les moyens cités pour travailler sur la valorisation des collections, nous relevons ceci :

> Un troisième média, permettant de toucher un public encore plus large est le média numérique, à savoir les sites internet des Archives de l'Etat et du CegeSoma, et des sites spécialisés comme Belgium WWII [32], Belgian War Press ou EHRI. Il s'agirait là de mettre en exergue, via de courtes info-fiches, des fonds spécifiques conservés au CegeSoma et dans les divers dépôts des Ar-

31. Renommé par ailleurs « accompagnement du public dans les collections » (Maerten, 2019, p. 1).

32. Lancée en 2017, cette plateforme a pour objectif de *fournir des informations sommaires mais essentielles sur l'histoire de la Belgique avant, pendant et après la Seconde Guerre mondiale* (CegeSoma, 2018b, p. 24), voir : https://www.belgiumwwii.be/.

> chives de l'Etat relatifs aux deux guerres mondiales, à leurs antécédents et à leurs conséquences. (Maerten, 2019, p. 6-7).

L'auteur mentionne également une dimension peu soulignée jusque là :

> Notons encore que si l'on veut vraiment aider le public scientifique ou autre à explorer les sources adéquates pour l'étude d'un thème, il serait nécessaire de l'aiguiller aussi vers d'autres centres d'archives en Belgique et à l'étranger. L'exemple bien connu de nous du thème de la Résistance le montre à souhait. (Maerten, 2019, p. 3).

À l'instar de la valorisation, l'histoire publique et la recherche fondamentale ne peuvent être pensées indépendamment de l'accès numérique aux collections. Ainsi, bien que les *digital humanities* soient seulement mentionnées à la pénultième page de la note portant sur la recherche fondamentale, son auteur, Dirk Luyten, explique qu'il s'agit d'un axe de recherche transversal pouvant être appliqué indifféremment à chacun des cinq thèmes-clés de recherche présentés au cours des pages précédentes :

> De uitdaging erin bestaat te opteren voor een onderzoeksmethode die een meerwaarde biedt vanuit het perspectief van het historisch onderzoek én homogene en kwalitatief hoogstaande bronnencorpora samen te stellen. Dit domein is nog in volle ontwikkeling, maar gedacht kan worden aan analyse van juridische en politieke discours, vormen van datamining om serieel onderzoek te faciliteren, of nog het compileren van specifieke bronnencorpora voor life course onderzoek [33]. (Luyten, 2018, p. 21)

En outre, la dimension de coopération refait son apparition ici :

> Digital humanities leent zich uitstekend tot (interdisciplinaire) samenwerking met andere instellingen, ook internationaal waarbij bijvoorbeeld corpora van gedigitaliseerde clandestiene pers uit verschillende landen kunnen gebruikt worden om transnationaal aspecten van de bezettingsgeschiedenis te onderzoeken [34]. (Luyten, 2018, p. 21)

33. [Traduction libre] L'enjeu est d'opter pour une méthode de recherche qui offre une valeur ajoutée du point de vue de la recherche historique et la constitution de corpus sources homogènes et de qualité. Ce domaine est encore en plein développement, mais on pourrait penser à l'analyse du discours juridique et politique, à des formes d'exploration de données pour faciliter la recherche en série, ou à la compilation de corpus de sources spécifiques pour la recherche sur les parcours de vie.

34. [Traduction libre] Les humanités numériques se prêtent parfaitement à une coopération (interdisciplinaire) avec d'autres institutions, également internationales, grâce auxquelles, par exemple, des corpus de presse clandestine numérisée de différents pays peuvent être utilisés pour étudier les aspects transnationaux de l'histoire de l'occupation.

Cette dimension internationale pourrait d'ailleurs être mise en relation avec le positionnement stratégique de l'institution relevé plus tôt dans cette note :

> Het CegeSoma heeft wetenschappelijke expertise ontwikkeld op een aantal onderzoeksdomeinen rond de beide wereldoorlogen, is nationaal en internationaal gekend en fungeert in België als een *contactpunt* voor (een deel van) het onderzoek over de beide wereldoorlogen en de herinnering. [...] Het CegeSoma wil zijn rol als *draaischijf* tussen onderzoekers uit Vlaanderen en Franstalig België en tussen België en het buitenland verder blijven spelen. Contacten met het buitenland zijn van belang om de band met het internationale onderzoek niet te verliezen en aansluiting te vinden bij nieuwe ontwikkelingen in het onderzoek[35]. (Luyten, 2018, p. 2)

Commme l'écrit Chantal Kesteloot – responsable de l'histoire publique au CegeSoma –, « le vaste chantier des humanités numériques concerne également l'histoire publique » et s'avère indispensable « si l'on souhaite être présent par-delà l'espace belge » (Kesteloot, 2019, p. 15). Citant la nécessité de développer des sites préexistants tels que Belgium WWII ou War Press, elle prévient toutefois :

> Nous disposons des connaissances, des collections mais nous manquons à la fois de ressources humaines et de compétences sur le plan numérique. (Kesteloot, 2019, p. 15)

Ainsi, le projet Belgium WWII, après avoir bénéficé de sources de financements externes ayant permis d'engager deux historiennes, ne dispose par exemple plus aujourd'hui d'aucun financement externe et repose entièrement sur l'équipe permanente du CegeSoma, ce qui s'avère particulièrement lourd en raison de la dimension multilingue du site[36].

Si ces différents documents ont été réalisés de façon isolée et aux fins d'un usage interne, il est toutefois intéressant de relever comment le numérique s'y retrouve de façon transversale, et ce alors même que le sujet est considéré comme un sujet à part, comme en témoignent l'attribution de cette charge à l'une des membres du personnel scientifique, de même que

35. [Traduction libre] Le CegeSoma a développé une expertise scientifique dans un certain nombre de domaines de recherche autour des deux guerres mondiales, est connu au niveau national et international et agit en Belgique en tant que point de contact pour (une partie de) la recherche sur les deux guerres mondiales et ses aspects mémoriels. Le CegeSoma souhaite continuer à jouer son rôle de plaque tournante entre les chercheurs belges flamands et francophones et entre la Belgique et l'étranger. Les contacts avec les pays étrangers sont importants pour ne pas perdre le lien avec la recherche internationale et rester connectés aux nouveaux développements en matière de recherche.

36. Elle relève ainsi : « en l'absence d'un *alter ego* néerlandophone, dans les faits, ce bilinguisme est devenu un frein à la mise en ligne de contenus » (Kesteloot, 2019, p. 13).

l'arborescence des dossiers communs sur le serveur du Centre (il y a, d'une part, la « gestion des collections » et, d'autre part, le « soutien numérique »). Par ailleurs, et cela s'explique probablement par la prédominance d'historiens (plutôt que d'archivistes) au sein du personnel scientifique, l'absence d'attention portée à l'évolution des standards archivistiques et notamment à la transition initiée par la publication du modèle conceptuel Records in Contexts semble significative et révélatrice d'un faible intérêt pour ce sujet. Pourtant, ce sont peut-être justement les données numériques – et plus précisément les données numériques lisibles par des machines, au cœur des nouveaux modèles conceptuels tels que Records in Contexts – qui pourraient permettre au Centre de se rapprocher des différentes ambitions dévoilées au cours des paragraphes précédents.

En effet, comme le recommande la responsable de l'histoire publique dans sa note, l'une des priorités pour le Centre devrait être de « penser les initiatives sous forme de projet *total* [...] compte tenu des moyens humains limités » (Kesteloot, 2019, p. 18). Or, les données constituent un dénominateur commun entre les différents secteurs de l'institution. Elles ont le potentiel d'amorcer une forme de travail en synergie à même de dépasser les pratiques cloisonnées du Centre. Elles constituent en effet la *matière première*, au carrefour entre accès aux collections, valorisation, histoire publique, recherche, projets en humanités numériques, nouvelles formes de collaboration et rayonnement international.

Si nous revenons à notre sujet d'étude et considérons en particulier les données d'autorité relatives à des personnes, telles que les données relatives à des figures engagées dans la Résistance belge au cours de la Seconde Guerre mondiale, il s'avère qu'elles constituent un point d'entrée vers les collections du Centre ; qu'elles servent de base pour mettre en exergue des personnalités marquantes dans le cadre de projets d'histoire publique ; qu'elles sont susceptibles de faire l'objet d'investigations plus poussées dans le cadre de projets de recherche ; qu'elles peuvent être réutilisées dans le cadre de projets en humanités numériques ; et enfin, qu'elles représentent des traits d'union entre les fonds d'archives du Centre et des ressources externes.

Cependant, le fait que de telles données existent ne suffit pas. En effet, comme nous l'avons vu précédemment, il s'agit de mettre à disposition des données correctes, normalisées et interopérables (Gillet, 2019). De plus, il s'agit également de pouvoir y accéder de façon aisée. Or, comme nous le verrons au cours du chapitre suivant, les données du CegeSoma sur des personnes constituent pour l'heure davantage une série d'îlots isolés qu'une véritable base de données centralisée. Ce constat nous a ainsi amenée à entendre à plusieurs reprises et de la part de différents membres du personnel qu'il serait véritablement utile pour eux de pouvoir bénéficier d'un accès cen-

tralisé à ces données, après qu'elles aient pu être dédupliquées, de manière à pouvoir retrouver une personne d'un clic.

3.2.2 Besoins des utilisateurs

Afin de pouvoir identifier les besoins des utilisateurs, plusieurs ressources ont été utilisées : un premier volet est basé sur des informations et constats issus de sources disponibles en interne ou publiées au cours des dernières années, tandis qu'un second volet est basé sur les requêtes d'utilisateurs effectuées dans le catalogue en ligne Pallas. Mais avant cela, ajoutons un peu de contexte en précisant que le personnel du Centre décrit son audience, en 2015, comme une audience spécialisée comprenant à la fois des étudiants et des chercheurs (Belges ou internationaux) issus de disciplines comme l'histoire, l'histoire de l'art, et les sciences de l'information ; des journalistes ; des éditeurs ; des commissaires d'exposition ; ainsi que des historiens intéressés par l'histoire locale ou familiale. Le personnel estime que ces différents publics correspondent à son audience cible, bien que le grand public avec un intérêt dans les conflits du XX^{e} siècle pourrait toutefois être mieux atteint (Hungenaert, 2016). Enfin, plus récemment, dans une note sur l'*Accompagnement du public dans les collections* (Maerten, 2019), Fabrice Maerten – responsable de la valorisation des Collections – identifie six types de publics auxquels le Centre devrait porter d'avantage d'attention :

— les cercles d'histoire locale
— les membres de familles de collaborateurs ou résistants
— les millieux universitaires (belges, mais aussi de pays limitrophes ou liés à l'histoire de la Belgique)
— les enseignants en histoire
— les musées et centres de documentation, les Maisons de la mémoire
— les médias.

Il relève par ailleurs les thèmes liés à la Seconde Guerre mondiale [37] suscitant le plus d'intérêt :

> Le public aime particulièrement se pencher sur l'histoire militaire, l'histoire locale et surtout l'histoire personnelle de membres de l'entourage familial. Les engagements surtout, notamment dans la résistance ou dans la collaboration, éveillent toujours une vive curiosité. (Maerten, 2019, p. 3)

Ces premiers éléments ayant été exposés, nous pouvons maintenant nous pencher sur les besoins des utilisateurs. L'analyse de ces besoins est basée sur les résultats issus du projet MADDLAIN [38] auquel nous avons activement pris

37. L'un des principaux domaines d'expertise du Centre.
38. Ce projet BRAIN, financé par la Politique scientifique fédérale belge, s'est déroulé de 2015 à 2017 et visait à fournir des données sur les pratiques et les besoins des utilisateurs

part. Il s'est construit autour de plusieurs volets d'analyse : des entretiens réalisés auprès du personnel des institutions ; une enquête en ligne ayant permis de collecter les avis de plus de 2000 répondants ; le suivi des traces d'usage sur les sites et catalogues en ligne des trois institutions ; ainsi que deux volets annexes, dédiés au *e-learning* et aux environnements virtuels de recherche. Les prochains paragraphes passent en revue une série de constats et de remarques tirés des rapports de ce projet, ainsi que d'un compte-rendu issu d'une journée d'ateliers organisée en février 2018 dans la continuité du projet MADDLAIN : *Le numérique aux Archives de l'État pour répondre aux besoins des Universités* (Depoortere *et al.*, 2018). Ces constats sont regroupés autour de trois thèmes : les données et métadonnées ; les voies d'accès aux collections ; l'édition collaborative.

Premièrement, au sujet des données et métadonnées, les participants à l'enquête en ligne pointaient du doigt un manque de communication sur ce qui est disponible en ligne, ce qui est disponible seulement sur papier, ce qui est déjà numérisé, dans quelle langue, etc. (Hungenaert et Gillet, 2017). Deuxièmement, ils étaient critiques concernant la qualité des métadonnées descriptives, déclarant qu'ils faisaient souvent face à des descriptions inexactes ou incomplètes. Ce que corroborent les remarques d'un panel de chercheurs, qui mettaient en exergue les problèmes de qualité des mots-clés, générateurs de bruit et de silence lors de l'affichage des résultats (Paul, 2017). Troisièmement, les utilisateurs déploraient que certains documents soient décrits seulement en français ou en néerlandais, restreignant ainsi le nombre de résultats affichés, notammment pour plus de la moitié des utilisateurs ayant déclaré ne jamais taper leurs mots-clés dans une autre langue que leur langue maternelle (Hungenaert et Gillet, 2017). Ces trois dimensions (communication, qualité, multilinguisme) étaient également présentes parmi les constats des chercheurs et universités : ils s'accordaient sur l'impossibilité de régler tous les problèmes de traduction et d'incohérence, estimant qu'il faudrait passer, d'une part, par une communication claire sur la question du multilinguisme et de la synonymie, et, d'autre part, par le développement de projets liés aux *Linked Data* (Depoortere *et al.*, 2018), une piste également évoquée dans le rapport final du projet MADDLAIN comme solution pour pouvoir afficher les mêmes résultats, indépendamment de la langue utilisée pour interroger la base de données (Hungenaert et Gillet, 2017).

En ce qui concerne les voies d'accès aux collections : tout d'abord, les participants à l'enquête MADDLAIN ont décrit les trois catalogues en ligne

de trois établissement scientifiques fédéraux, dans l'optique de moderniser l'accès à leurs données numériques ; l'ensemble des résultats du projet est disponible sur le site web du CegeSoma : https://www.cegesoma.be/fr/project/le-projet-maddlain

comme étant peu *user-friendly*, en raison de leur structure compliquée et difficile à utiliser (Hungenaert et Gillet, 2017). Deuxièmement, si l'analyse des traces d'usage a montré que les modes de recherche avancée n'étaient utilisés que par une minorité (Chardonnens, 2017), il est intéressant de noter que le panel des chercheurs y accorde de l'importance. En effet, ces derniers ont exprimé une volonté de pouvoir bénéficier de plusieurs portes d'entrée et donc de plusieurs méthodes de recherche, en mixant des recherches intuitives de type Google sur base de mots-clés d'abord, et d'autres méthodes plus précises ensuite afin d'affiner leur recherche à l'aide de champs spécifiques (Paul, 2017). Par ailleurs, alors qu'ils étaient interrogé sur l'utilité de futurs outils ou sur des environnements virtuels de recherche, ils déclaraient être réticents si la courbe d'apprentissage était élevée, et surtout, ils insistaient sur le fait qu'un accès optimal aux collections était prioritaire (Paul, 2017). Même insistance du côté des représentants des universités, estimant que :

> La mission première des Archives de l'État est de rendre un maximum de fonds accessibles à la recherche, d'en assurer la structure et la description avec des métadonnées stables et cohérentes pour pouvoir les utiliser dans leurs propres environnements de recherche virtuels. (Depoortere *et al.*, 2018, p. 11)

Ils notent ainsi que « la multiplicité des outils de recherche et de bases de données ne favorise pas une vue d'ensemble » et qu'« il serait bienvenu d'intégrer les résultats au sein d'un seul catalogue général » (Depoortere *et al.*, 2018, p. 12). Cet accès aux données concerne tant les descriptions archivistiques que d'autres jeux de données que possèdent les Archives de l'État :

> Étant donné la diversité de sources connexes que conserve l'institution, la création de banques de données croisant ces différentes informations serait une réelle plus-value et un énorme gain de temps pour les chercheurs. (Depoortere *et al.*, 2018, p. 13)

Enfin, en ce qui concerne les projets d'édition collaborative – par le biais du *crowdsourcing* –, la moitié des répondants à l'enquête se déclarait prête à y participer (Hungenaert et Gillet, 2017). Quant aux partenaires universitaires, ils semblaient majoritairement ouverts à ce type de pratique, tout en précisant que l'idéal serait d'envisager un modèle *win-win* entre chercheurs et institutions, notamment en valorisant davantage les données brutes issues de recherches universitaires sur les collections des institutions, qui sont actuellement sous-exploitées. Ils soulignaient également que, s'ils sont réceptifs à l'idée de valoriser le résultat de leurs projets ou de contribuer occasionnellement à des plateformes, l'objectif devrait selon eux être :

> La création de bases de données interopérables, permettant de créer des liens vers d'autres plateformes externes à l'institution

> (Wikipedia, Wikidata, Wikisource, autres Linked Open Data, plateformes d'institutions partenaires, ...). (Depoortere *et al.*, 2018, p. 13)

À ces observations plus générales s'ajoutent des considérations portant plus spécifiquement sur les besoins liés à la recherche de personnes. Ce second volet d'observations est basé sur les requêtes d'utilisateurs effectuées au sein du catalogue en ligne Pallas. En effet, comme le relevait Bearman aux sujets des requêtes d'utilisateurs et points d'accès :

> Logically, user queries should be the point of departure for defining a strategy to augment access. Incredibly, archivists have no published literature of user-queries analysis with which to begin. (Bearman, 1989)

À notre connaissance, ce constat est toujours d'actualité et nous n'avons pas rencontré ce type d'analyse au cours de notre revue de la littérature. Or, notre participation active au projet MADDLAIN – comme évoqué précédemment, ce projet de recherche, fruit d'une collaboration entre le CegeSoma, les Archives de l'État et KBR, visait à étudier les pratiques et besoins numériques de leurs utilisateurs – nous a précisément permis d'avoir accès aux requêtes effectuées dans les catalogues en ligne de trois établissements scientifiques fédéraux, parmi lesquels figure Pallas, le catalogue du CegeSoma. Nous avons ainsi mené une analyse visant à connaître le ratio de requêtes comportant des entités nommées, et plus précisément des noms de personnes.

Pour des raisons de clarté et de concision, nous avons choisi de placer le détail de cette analyse en annexes. Ainsi, l'Annexe 2 [39], présente de façon exhaustive les quatre étapes ayant permis de traiter et d'étudier ces requêtes : la collecte des données, le pré-traitement des données brutes, l'extraction de noms de personnes et enfin la réconciliation de ces noms à l'aide de bases de connaissance externes. En résumé, ces étapes nous ont permis de passer d'un jeu de données couvrant une période d'un an et contenant 30 703 requêtes normalisées [40], à un sous-ensemble contenant 21 385 requêtes distinctes [41]. Parmi ces plus de 20 000 requêtes, un script d'extraction de (potentiels) noms de personnes nous a permis d'isoler un sous-ensemble de 2885 requêtes distinctes. En tenant compte de la fréquence où ces (possibles) noms

39. Page 323.

40. Pour des détails sur ce que ce terme recouvre dans le contexte de cette analyse, consulter la sous-section de l'Annexe 2 dédiée au pré-traitement des données, page 325.

41. Il s'agit de compter le nombre de termes distincts, indépendamment du fait que la requête ait pu être formulée à plusieurs reprises. Ainsi, si trois personnes différentes ont effectué une requête sur la résistante *Marguerite Bervoets*, il s'agit de ne compter la requête qu'une seule fois.

de personnes ont été utilisés, nous arrivons à un total de 4 104. Cette première estimation signifie donc qu'au cours de la période d'analyse, plus d'une requête sur dix [42] sur le catalogue en ligne du CegeSoma contenait un (probable) nom de personne.

Cette estimation est toutefois à considérer avec précaution. En effet, ce sous-ensemble composé de 4 104 possibles noms de personnes est susceptible de contenir de faux positifs, tels que *Happy Birthday* : en effet, bien que Happy constitue un véritable prénom – présent dans la liste de prénoms de référence utilisé par l'algorithme d'extraction –, sa présence dans cette requête ne semble pas faire référence ici à un prénom, mais à une expression en anglais relative à un anniversaire. Ainsi, pour affiner la pertinence de ce sous-ensemble, nous avons opté pour la réconciliation de ces potentiels noms de personnes avec des personnes présentes dans des bases de connaissance externes. Cette étape de réconciliation [43] avec les données de Wikidata et de Viaf nous a permis de restreindre le périmètre initial de 2 885 (potentiels) noms distincts à un sous-ensemble de 1 541 noms ayant pu être associés à des entités Wikidata et, ou VIAF. La figure 2 montre un aperçu de la distribution de ces réconciliations, de même que la complémentarité des deux sources externes utilisées au cours de cette étape.

Si cette dernière étape visant à affiner et rendre plus rigoureuse la reconnaissance de noms de personnes [44] pourrait être poussée plus loin encore, en explorant par exemple le profil des personnes recherchées à l'aide de propriétés Wikidata comme la profession, la nationalité ou la participation à des conflits de ces individus, elle permet déjà de quantifier à l'aide de données empiriques la proportion de requêtes portant sur des noms de personnes. Concrètement, ces 1 541 noms réconciliés sont présents dans 2 502 requêtes distinctes sur un total 30 703, soit 8,15%. Ce pourcentage est par ailleurs susceptible d'avoir augmenté depuis 2015, étant donné qu'une hausse des demandes relatives à des personnes a été observée par le personnel du CegeSoma au cours des dernières années. Cela s'explique par différentes causes. Ainsi, nous pouvons lire dans le rapport annuel 2017 du Centre :

> La publication du guide « Papy était-il un nazi » [...] et les septs émissions de la série télévisée *Kinderen van de collaboratie* ont provoqué une forte hausse du nombre de questions, souvent de la part de proches se renseignant sur le passé de collaboration de certains membres de leurs familles. [...] [Outre] les demandes relatives à la consultation de dossiers judiciaires [...], les deux

42. 4 104 représente 13,36% du total de 30 703 requêtes.

43. Présentée de façon détaillée dans la sous-section de l'Annexe 2 dédiée à la réconciliation, page 328.

44. Qui reste toutefois limitée au contenu que couvre VIAF et Wikidata dans le cas présent.

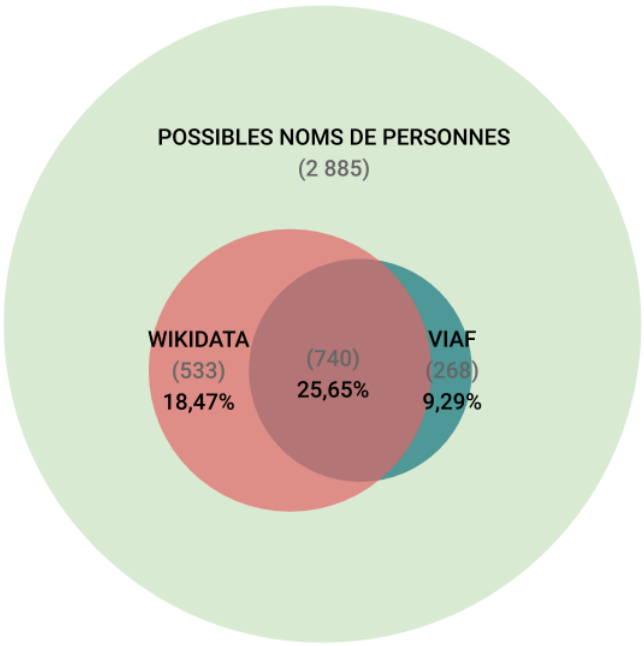


FIGURE 3.1 – Aperçu de la répartition des réconciliations des (possibles) noms de personnes avec VIAF et Wikidata (diagramme à l'échelle).

> autres thèmes centraux sur lesquels les questions sont les plus fréquentes concernent les activités de résistance, le travail en Allemagne et les prisonniers de guerre. (CegeSoma, 2018b, p. 12)

Ce phénomène s'est poursuivi avec la publication par le CegeSoma d'un guide sur les sources de la Résistance en Belgique, publié en 2020 sous le titre *Papy était-il un héros ?* Son auteur, Fabrice Maerten – responsable de la valorisation des collections au CegeSoma –, explique ainsi qu'il est principalement confronté à deux types de demandes : la majorité des questions proviennent de particuliers recherchant des informations sur des proches impliqués dans la Résistance et dès lors intéressés par tout ce qui concerne une personne en particulier ; tandis qu'un plus faible pourcentage des demandes provient de chercheurs ou historiens amateurs effectuant une recherche ciblée sur un groupe de personnes, comme par exemple les personnes engagées dans des activités de Résistance dans une certaine localité [45].

3.2.3 Analyse fonctionnelle

Ce chapitre se clôture avec une analyse fonctionnelle synthétisant les besoins relatifs à la gestion des données d'autorité du CegeSoma. Une liste d'exigences et priorités, issue de la synthèse des éléments évoqués au cours

45. Propos échangés au cours d'un rendez-vous de travail, 18.06.2019.

des pages précédentes, a été discutée et enrichie au cours d'un dialogue continu avec Florence Gillet, la responsable de la numérisation du Centre, de manière à pouvoir esquisser ce qui pourrait s'apparenter à un cahier des charges concernant la gestion des données d'autorité du CegeSoma. Comme nous le verrons, cette liste amorce une transition vers le logiciel libre Wikibase, pressenti comme une solution permettant de répondre à ces exigences et étant dès lors amené à jouer un rôle central dans le cadre de la gestion des données d'autorité du Centre.

Les besoins relevés et décrits au cours des pages précédentes ont été traduits en tâches et fonctionnalités regroupées au sein de huit sous-points : traitement des données ; modélisation ; implémentation ; édition manuelle ; importation automatisée ; administration et contrôle ; accès aux données ; modes de recherche.

1. **Traitement des données**
   - Standardisation (par exemple : dates, langues)
   - Sémantisation éventuelle lorsque c'est jugé pertinent (par exemple : désambiguïsation des noms de lieux ou professions en les associant à des URIs)
   - Déduplication : opération d'*entity linking* afin d'éviter les doublons
   - Alignement : création de liens entre les noms de personnes et des ressources externes (par exemple : Wikidata)
2. **Modélisation**
   - Modèle de données personnalisé : conception d'un modèle sur mesure permettant de prendre en charge des notions de provenance et d'incertitude
   - Interopérabilité : réutilisation maximale de vocabulaires préexistants (comme les propriétés Wikidata et, ou l'ontologie en cours de création par l'International Council on Archives : Records in Contexts Ontology[46]), afin de garantir la plus grande interopérabilité
   - Sources : maintien d'un lien entre les nouvelles entités et le fichier duquel elles sont issues
   - Incertitude : mise au point d'une pratique permettant de gérer une information incertaine (exemple : l'un des chiffres d'une date de naissance est illisible)
   - Données liées : possibilité d'intégrer des renvois vers des ressources externes
   - Sources : système permettant de spécifier au niveau de granularité le plus fin la provenance de l'information

46. https://www.ica.org/standards/RiC/ontology.

3. **Implémentation**
   — Installation : installation facilitée à travers un mécanisme de type Docker
   — Configuration : personnalisation de l'instance (exemples : apparence, logo, langue, extensions, ...)
4. **Édition manuelle**
   — Interface graphique : possibilité d'édition manuelle des données, y compris par des personnes ne possédant pas de connaissance informatique avancée (exemple : bénévoles âgés du Centre)
   — Édition simultanée : possibilité d'édition collaborative et simultanée de données décrivant une même entité
   — Multilinguisme : système prenant en compte la gestion de données multilingues (exemples : français, néerlandais, anglais, allemand)
   — Convivialité : fonctionnalité d'autocomplétion automatique des données au moment de l'encodage
   — Doublon : fonctionnalité d'alerte en cas de création de doublon
   — Formulaire : outil permettant de générer des formulaires affichant les propriétés devant être encodées pour un certain type d'éléments (exemple : propriétés devant systématiquement être remplies pour une personne)
   — Contraintes : possibilité d'établir des contraintes spécifiant le type de valeurs utilisées pour une certaine propriété (exemple : dates, coordonnées géographiques, quantités, ...)
5. **Importation automatisée**
   — Création de données : système permettant le chargement automatisé de jeux de données (exemples : fichiers CSV) sous forme de lots
   — Modification de données : système permettant la modification en masse des données contenues dans l'instance
6. **Administration et contrôle**
   — Permissions : possibilité d'attribuer des droits distincts en fonction du ou des rôles (non exclusifs) attribués aux utilisateurs (exemples : lecteur [47] ; éditeur [48] ; responsable du contrôle qualité [49] ; administrateur [50] ; robot [51])

47. Peut seulement consulter les données.
48. Peut créer, compléter et modifier des éléments et des pages.
49. Peut fusionner deux éléments ; supprimer des éléments ; créer, modifier ou supprimer des propriétés.
50. Peut bloquer des utilisateurs ; protéger ou supprimer des pages.
51. Peut effectuer des modifications massives à un rythme soutenu.

   - Suivi : système permettant un suivi personnalisé des modifications de données
   - Historique : enregistrement systématique de toutes les opérations effectuées sur les données

7. **Accès aux données**
   - Affichage : interface multilingue
   - Export : possibilité d'exportation de données dans différents formats incluant au minimum le format CSV
   - API : possibilité d'utiliser une API pour interroger l'instance à l'aide d'applications tierces
   - Dump : fonctionnalité d'export de l'ensemble des données (dump)
8. **Recherches**
   - Recherche simple : possibilité de recherche multilingue, avec autocomplétion des termes de requête et prise en compte des formes alternatives d'une entité
   - Requêtes structurées : possibilité de formuler des requêtes complexes tirant parti du caractère structuré des données encodées, en particulier grâce au langage SPARQL

Cette analyse fonctionnelle a permis de confirmer le choix pressenti de Wikibase pour créer une plateforme de coproduction de données d'autorité pour le CegeSoma. Ce choix est principalement motivé par son caractère *open source*, par le fait qu'il est en développement continu, ainsi qu'en raison des nombreuses fonctionnalités qu'il offre, telles que la création d'une ontologie sur mesure, l'attribution d'URIs, l'édition collaborative de données structurées multilingues, et l'exportation de données dans divers formats. Cette plateforme, prenant la forme d'un entrepôt de données s'appuyant sur les technologies du Web sémantique, est destinée à accueillir en premier lieu [52] un référentiel Personnes physiques [53] du CegeSoma. Le chapitre suivant expose en détails les divers jeux de données concernés, de même que la façon dont ils vont être préparés et modélisés avant leur importation dans une instance Wikibase.

52. À terme, elle doit être en mesure d'inclure d'autres données, comme d'autres types de référentiels (organisations, lieux, sujets, ...) susceptibles d'être publiés dans les années à venir par le CegeSoma ou les Archives de l'État.

53. Producteurs d'archives, mais également des noms de personnes issus de l'indexation-matière.

4 | Données

Introduction

Dans ce nouveau chapitre, nous présentons les différents corpus de données empiriques au cœur de notre étude de cas, nous détaillons les étapes de traitement précédant leur intégration dans une instance Wikibase dédiée et enfin nous discutons de la modélisation des données.

La première section introduit les données du CegeSoma sous la forme d'un continuuum allant des données les moins structurées au plus structurées. Elle décrit ensuite l'échantillon utilisé dans le cadre de notre étude de cas, ainsi que les considérations ayant dû être prises en compte en ce qui concerne le respect de la vie privée.

La deuxième section décrit le travail effectué sur cet échantillon de données, à savoir : pré-traitement, *entity linking*, et réconciliation avec des identifiants externes. Nous illustrons ainsi comment ces étapes aboutissent à la création d'une nouvelle *couche* sur les noms de personnes présents dans cet échantillon, en les reliant à des référentiels externes. En effet, les URIs issus de la réconciliation viennent s'ajouter aux métadonnées préexistantes et peuvent ensuite être utilisés comme pivots facilitant l'interconnexion vers les documents ou bases de données provenant d'autres institutions.

Enfin, la troisième section révèle comment ces données nettoyées, dédupliquées, réconciliées et potentiellement enrichies sont adaptées au modèle de données Wikibase. Cette section détaille comment sont modélisées les données d'identification, les propriétés propres au champ de la Seconde Guerre mondiale, ainsi que les relations à d'autres personnes ou d'autres ressources.

4.1 Corpus

4.1.1 Continuum

Cette première section vise à présenter de façon exhaustive le corpus au cœur de cette étude de cas. En effet, bien que seul un échantillon de don-

nées (voir 4.1.2) soit utilisé dans le cadre de notre prototype, nous souhaitons mettre l'accent sur le caractère hétérogène des données à prendre en considération. En effet, dans les faits, il s'est avéré que le Centre ne dispose pas d'un fichier d'autorité unique pour les personnes physiques liées à ces collections. Ainsi, si le système de gestion documentaire du CegeSoma (Pallas) dispose, conformément à la tradition bibliothéconomique, de deux listes d'autorité distinctes pour les personnes physiques [1], il est apparu que des noms de personnes sont également présents au sein d'autres types de documents ou de métadonnées, qui seront décrits au cours des paragraphes suivants. Or, si nous envisageons tous les noms de personnes comme des chaînes de caractères pouvant potentiellement être liées à une notice d'autorité d'un référentiel unique dévolu aux entités Personne – elles-mêmes identifiées à l'aide d'URIs représentant des personnes du monde réel –, il devient dès lors nécessaire de pouvoir lier ces diverses mentions de noms de personne, quel que soit le support sur lequel elles sont mentionnées. Cela signifie l'abolition de la frontière séparant données et métadonnées : nous incluons dans notre vision à la fois les noms présents dans les documents eux-mêmes – par exemple des entités nommées rencontrées parmi des corpus de presse ancienne numérisée – et les noms présents dans des métadonnées descriptives [2]. Dans le cadre de cette étude de cas, nous utilisons donc le terme *données* en l'envisageant dans son acception la plus large : c'est-à-dire en incluant de façon indifférenciée données et métadonnées.

Comme l'a montré Boydens, il arrive que coexistent des informations de nature hétérogène au sein d'un même système d'information, conduisant à la mobilisation des notions d'information structurée, non structurée et semi-structurée (Boydens, 2001). L'exploration des différentes listes et ressources du CegeSoma contenant des noms de personnes et l'observation du fait qu'il n'existe pas toujours de frontière nette entre ces notions, nous ont amenée à plutôt les envisager sous la forme d'un continuum. Un continuum qui serait constitué d'un spectre de données allant des moins structurées au plus structurées. La figure 4.1 offre une vue d'ensemble de ce continuum, dans lequel nous distinguons cinq types de données. À chacun de ces types, dont les particularités sont décrites au cours des prochains paragraphes, sont associés des jeux de données du CegeSoma [3].

1. Une liste d'autorité englobant auteurs, éditeurs et producteurs, ainsi qu'une liste de vedettes-matières incluant notamment des noms propres.

2. Leur traitement diffère cependant dans la mesure où il faut s'attendre à ce que les premiers comportent un degré d'ambiguïté élevé (par exemple si seuls les noms et prénoms sont présents dans le texte, sans être systématiquement accompagnés de données d'identification comme des dates de naissance et de décès), là où les seconds seront potentiellement mieux identifiés.

3. Précisons que de nouveaux jeux de données sont actuellement en cours de création et pourraient venir compléter cet aperçu qui reprend l'existant en date de la fin décembre

À l'extrémité gauche du continuum, se trouvent les données les moins structurées : c'est-à-dire du **plein texte** : des chaînes de caractères présentes au milieu d'un document analogique ou numérique – qu'il s'agisse d'un titre, d'un contenu textuel ou d'une légende –, sans qu'aucune balise ne signale qu'il s'agisse de noms de personnes. Rien n'indique formellement qu'il s'agit d'un nom de personne. Seul un œil humain ou un outil de reconnaissance d'entités nommées pourra identifier cette chaîne de caractères comme un (probable) nom de personne. Cela signifie que ces noms de personnes sont des données structurées en puissance : un processus manuel ou semi-automatisé d'extraction d'information pourrait par exemple permettre d'extraire et d'étiqueter certaines chaînes de caractères à l'aide d'une balise <personne>, mais pour l'instant, rien ne l'indique.

Vient ensuite une catégorie intitulée **tags**, c'est-à-dire des chaînes de caractères isolées, constituées de noms de personnes. Ils sont présents de façon plus structurée que dans le cas précédent : il peut s'agir de noms présents dans un texte qui auraient déjà été étiquetés en tant que personne à l'aide d'une balise dédiée, mais cette catégorie inclut également des listes de noms présents par exemple dans un index onomastique ou associés à un document sous forme de mots-clés. Les données sont donc présentes de façon semi-structurée : il arrive en effet que l'on retrouve au sein d'un même *tag* plusieurs données distinctes : nom de famille, initiales ou prénom(s), mais également d'autres informations comme un pseudonyme, des dates de naissance ou de mort.

Au centre du continuum, nous trouvons les **données tabulaires**. Cette catégorie contient des données stockées dans un tableau et structurées à l'aide de plusieurs colonnes. Cela signifie que contrairement à la catégorie précédente où tout était présent pêle-mêle dans un seul champ, ici l'information est décomposée en éléments distincts : nom, prénom, date de naissance, et autres renseignements portant sur une personne sont encodés séparément. Cette structuration plus élaborée des données [4] a des implications directes sur la façon dont ces informations peuvent ensuite être exploitées. Par exemple, si un lieu de naissance a été stocké dans une colonne distincte, il sera ensuite beaucoup plus aisé de proposer un filtre des résultats basé sur cette information, ou encore de réaliser un projet de visualisation de don-

2018. En effet, le Centre a récemment décidé – dans le sillage de la publication d'un ouvrage spécialisé sur les sources de la résistance en Belgique (Maerten, 2020) – de mettre en valeur les milliers de dossiers personnels contenus dans divers fonds d'archives liés à la résistance en Belgique. Il s'agit de divers fichiers Excel principalement encodés par les bénévoles du Centre.

4. Notons également que, bien que ce ne soit pas le cas ici, le contenu des cellules pourrait également être formellement spécifié à l'aide de contraintes d'intégrité. Par exemple, Excel propose un système de validation des données, qui peut être basé sur un format (date, par exemple) ou encore sur une liste déroulante.

nées en affichant sur un fond de carte les lieux de naissance d'un groupe de personnes [5]. En revanche, il faut noter que l'information ne peut pas être hiérarchisée contrairement à ce que permet la prochaine catégorie.

5. Sans cette structuration en colonnes le processus serait plus laborieux : il faudrait commencer par *parser* un champ unique contenant plusieurs types d'informations, afin de pouvoir extraire les lieux de naissance dans une colonne distincte.

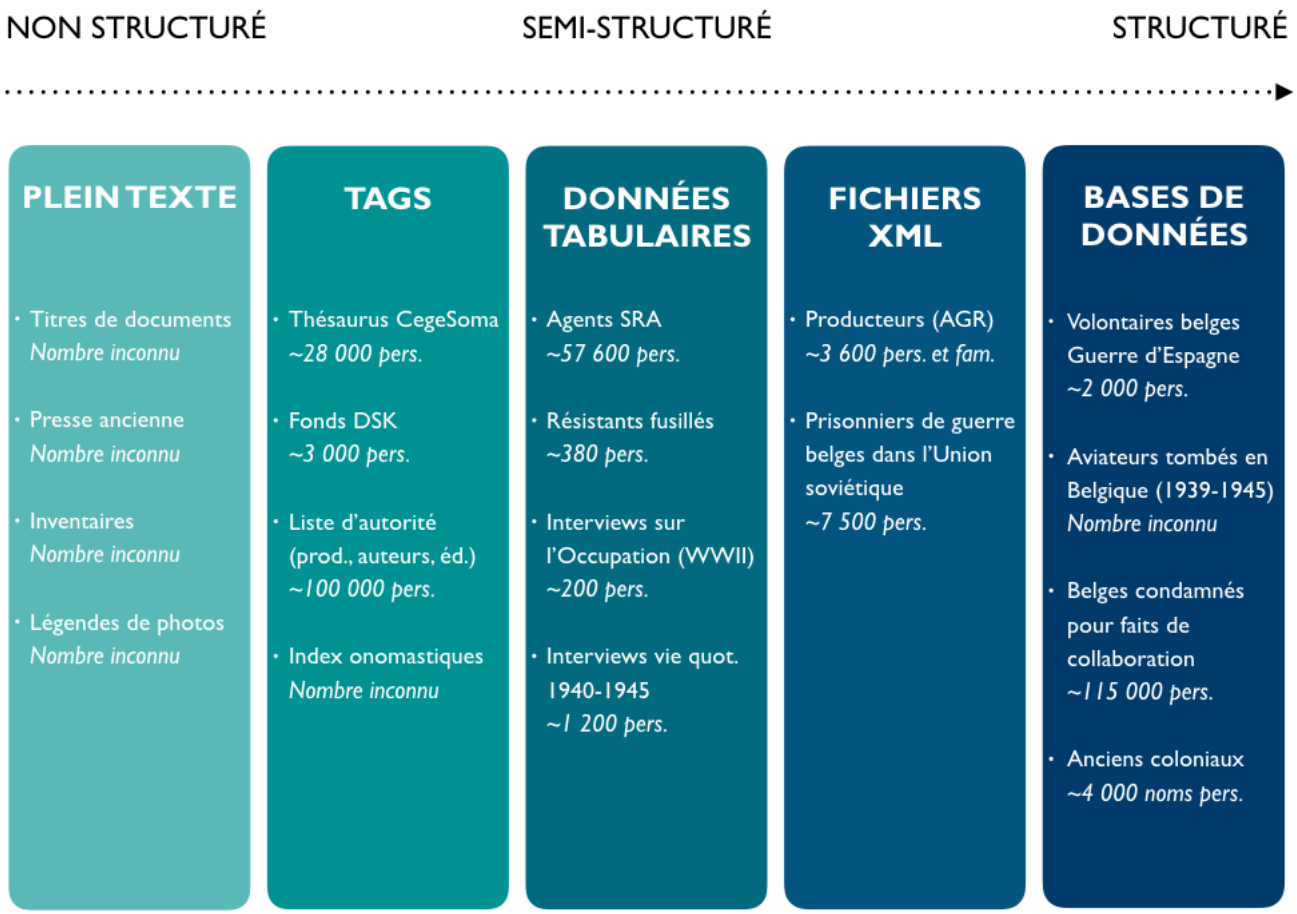


FIGURE 4.1 – Continuum des données nominatives du CegeSoma, des moins structurées aux plus structurées.

Le type suivant est dédié aux **fichiers XML**, qui sont basés sur l'*EXtensible Markup Language*. Ce langage informatique de balisage, personnalisable, a été créé afin de pouvoir échanger plus facilement des données. Les fichiers XML continuent de structurer chaque type d'information de façon distincte, comme les données tabulaires de la catégorie précédente, mais la structuration va plus loin encore. Cette fois-ci, ce sont des balises qui permettent de structurer les données dans un fichier texte. Ces balises encadrent un contenu, de la même façon que peuvent le faire les tags évoqués dans la première catégorie, mais avec davantage de complexité dans la mesure où elles permettent de structurer différents types de données de manière hiérarchisée et organisée. Ces balises permettent ainsi à des être humains de décrire des données de façon à ce qu'elles soient également lisibles par des machines. Les fichiers XML sont accompagnés de règles régissant leur structure et précisant l'imbrication des balises [6] : c'est ce caractère extrêmement structuré et hiérarchisé qui explique la place qu'occupent les fichiers XML dans le continuum que nous proposons.

À l'extrémité droite, désignant les données les plus structurées de notre continuum, nous avons placé les **bases de données** relationnelles. Ces dernières abritent des données stockées dans différentes tables, reliées entre elles à l'aide de *clés* uniques. Outre ces clés permettant d'éviter la création de doublons, d'autres contraintes d'intégrité sont utilisées afin de garantir la cohérence des données. Ces contraintes permettent de définir des règles visant par exemple à spécifier le format des valeurs acceptées dans un certain champ. Il faut souligner que le caractère structuré de l'information contenue dans ces tables peut ensuite être exploité à l'aide d'un langage informatique de requête reposant sur des opérateurs comme l'union, qui permet de combiner le contenu présent dans les différentes tables de la base de données.

Enfin, soulignons que ce continuum a été constitué au cours de l'observation des données que possède le CegeSoma. Il est clair que si priorité était donnée à une approche théorique plutôt qu'empirique, il faudrait inclure dans ce continuum une sixième catégorie incluant des **triplets RDF**. Ces données reposant sur des ontologies composées de classes, de propriétés et de contraintes sont en effet extrêmement structurées et destinées avant tout à être lisibles par des machines. Cependant, le CegeSoma n'ayant pas eu l'occasion par le passé de développer des projets allant dans ce sens, aucun élément de notre corpus ne justifie la présence de cette catégorie.

6. Ces règles sont formulées à l'aide de syntaxes comme DTD ou Schémas XML.

4.1.2 Échantillon

Si la figure 4.1 propose une vue récapitulative des ressources et fichiers dans lesquels nous retrouvons des noms de personnes, notre étude de cas se base en revanche sur un échantillon seulement. Ce choix s'explique par diverses raisons. Premièrement, certaines listes contiennent des données sensibles, à l'instar des listes nominatives basées sur les condamnations pour faits de collaboration ayant été publiées dans le journal officiel de l'État belge, le Moniteur belge. Ces listes n'incluent par ailleurs pas les dates de décès des personnes mentionnées. Ce type de contenu irait donc à l'encontre des mesures mises en place dans le cadre de cette thèse afin d'assurer le respect de la vie privée, mesures qui seront détaillées au cours de la prochaine sous-section. Deuxièmement, certaines données correspondent à des ressources papier n'ayant pas encore été numérisées et dont le traitement serait dès lors très laborieux, à l'instar d'index onomastiques contenus dans d'anciens instruments de recherche. Troisièmement, les données extrêmement incomplètes, limitées à des noms et prénoms – à l'instar d'une grande partie des producteurs d'archive [7]–, ne semblent pas idéales dans la mesure où leur intégration dans une Wikibase ne permet pas d'expérimenter pleinement les possibilités de modélisation de l'outil ; par ailleurs leur traitement devrait idéalement être accompagné de recherches complémentaires afin de pouvoir réduire l'ambiguïté associée à ces noms à l'aide par exemple d'années ou de lieux de naissance. Enfin, il faut garder à l'esprit que le traitement de ces jeux de données – dont les diverses étapes seront détaillées au cours de la section suivante – représente un processus chronophage nécessitant

7. Jusqu'il y a peu, les noms de producteurs étaient encodés au sein d'une liste d'autorité qui englobait également les auteurs et éditeurs et comptait 95 959 noms en octobre 2018. Cependant, avec la migration des descriptions des collections du CegeSoma vers SAM, le système de gestion des archives des Archives de l'État, les livres et manuscrits sont désormais destinés à être traités à l'aide d'un logiciel distinct. Cette liste d'autorité contenant pêle-mêle près de 100 000 noms a donc perdu de son intérêt, sachant qu'en outre seule une minorité d'entre eux correspondent en réalité à des producteurs d'archives. Par ailleurs, la perspective de la migration des données du CegeSoma a poussé le personnel du Centre a déployé de nouveaux efforts concernant le nettoyage des données. Ainsi, le récolement effectué en janvier 2019 fut notamment l'occasion d'effectuer une vérification des descriptions de près de 2 800 fonds d'archives du Centre. À cette occasion, un nouveau fichier récapitulatif a vu le jour, tenant compte du système de rubriques des Archives de l'État et permettant ainsi de filtrer les producteurs d'archives pour ne garder que les personnes [physiques] et les familles.

Ce jeu de données, contenant 1 150 noms distincts pour 1 293 fonds, est extrêment limité : à l'heure actuelle seuls les noms sont documentés. Dans le meilleur des cas, le nom de famille est accompagné d'un ou plusieurs prénom(s), parfois juste d'une initiale. De plus, il est apparu que le terme producteur a parfois été utilisé de manière très éloignée de ce que l'on entend au sens archivistique : en effet, certains fonds auraient pour producteurs le nom de la personne ayant versé le fond, alors que ce dernier avait été constitué par un autre membre de la famille ou encore une association pour laquelle travaillait cette personne. Un travail de nettoyage des données est donc requis pour (re)contextualiser ces fonds (Desmet, 2019, p. 32).

d'importantes ressources. Pour toutes ces raisons, nous avons décidé d'opter pour une stratégie pragmatique basée sur la sélection d'un échantillon permettant de tester l'étendue des possibilités de Wikibase, tout en limitant le nombre de données à manipuler. Nous présentons ces choix au cours des prochains paragraphes.

Premièrement, notre échantillon est constitué de l'un des principaux jeux de données du corpus, à savoir l'ensemble de noms propres issus du thésaurus du CegeSoma – que nous désignerons en tant que *thésaurus Pallas* au cours des pages suivantes. Ces noms propres étant mélangés aux autres descripteurs de façon indifférenciée, une première mise en contexte s'impose afin de découvrir l'historique et les tenants et aboutissants de ce langage d'indexation-matière.

Tout commence dans les années 1990, lorsque l'institution décide de se doter d'une liste de vocabulaire contrôlée. Cette dernière doit servir à décrire les collections du Centre, notamment photographiques, afin d'instaurer un accès thématique général. Le Centre crée donc en 1994 une première liste d'autorité, présentée sous forme de thésaurus : « Thesaurus de l'histoire de la Seconde Guerre Mondiale en Belgique (et de la période 1930-1950) ». Le vocabulaire contrôlé français RAMEAU (Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié) sert de base à la création de ce thésaurus. Étant donné la spécificité des collections du Centre et l'emphase mise sur la Seconde Guerre mondiale, les vedettes RAMEAU sont complétées à l'aide des mots-clés ayant été utilisés jusque-là pour décrire les collections du Centre (Temmermann et Waeyenberg, 2000). Ce thésaurus continue d'évoluer lorsque le Centre se transforme en 1997 et devient l'actuel Centre d'Études et de Documentation Guerre et Sociétés contemporaines. La réorientation de son périmètre recherche, qui s'étend désormais à l'ensemble du XX^{e} siècle et prend une dimension internationale, signifie que la première version du thésaurus Pallas n'est plus suffisante et doit être complétée [8], donnant lieu à la publication d'une nouvelle édition, en l'an 2000 (Temmermann et Waeyenberg, 2000). Enfin, vu le caractère bilingue de l'institution, tous les mots-clés du thésaurus sont systématiquement traduits en français ou néerlandais.

Avant de poursuivre, il faut souligner que si ses auteurs – Patrick Temmerman et Stéphanie Waeyenberg – le qualifient de *thesaurus*, il ne s'agit

8. Il s'agit d'inclure les descripteurs RAMEAU couvrant les domaines culturel, militaire, politique et socio-économique afin de pouvoir refléter les grandes lignes de l'histoire internationale du début du XX^{e} siècle jusqu'à nos jours. De plus, de nouveaux mots-clés – le terme vedette-matière ou descripteur semble plus adéquat, néanmoins, le personnel du Centre et l'interface de recherche en ligne privilégiant le terme *mot-clé*, nous l'utilisons également dans le cadre de ces pages – sont également créés par le Centre afin de couvrir l'histoire de la Belgique en particulier, en suivant les règles de syntaxe RAMEAU.

toutefois pas d'un thésaurus au sens strict. En effet, l'une des grandes différences entre liste d'autorité et thésaurus repose sur le caractère pré-coordonné du premier – les termes sont associés au cours du catalogage –, par opposition au caractère post-coordonné du second – l'utilisateur associe des termes au moment de sa recherche. En outre, un thesaurus est conçu a priori, tandis qu'une liste d'autorité telle que RAMEAU évolue au fur et à mesure des besoins documentaires et des propositions formulées par les utilisateurs eux-mêmes. Le langage d'indexation du CegeSoma s'apparente donc davantage à une « liste d'autorité de mots-clés précoordinatifs dans laquelle des relations ont été établies sur le modèle du thesaurus » (Temmermann et Waeyenberg, 2000). Concrètement, ces mots-clés sont construits selon une syntaxe bien déterminée et reliés à d'autres par une ou plusieurs relations (génériques, associatives ou d'équivalence). Cela signifie qu'à cette période, les responsables de l'indexation des collections se référaient à la version publiée du thésaurus afin d'identifier les descripteurs précoordonnés les plus adaptés à leurs besoins [9].

Avec le temps, le Centre s'est peu à peu retrouvé privé des ressources humaines requises pour maintenir ce thésaurus à jour et ajouter de nouvelles entrées tout en ayant le souci de préserver sa qualité [10]. Ou du moins, personne n'a eu le souci de mettre en place une politique documentaire stricte à ce sujet et d'affecter du personnel à cette tâche. Concrètement, cela se traduit par le fait qu'aucune nouvelle édition n'a vu le jour depuis 20 ans. Pour autant, cela ne signifie pas que nul nouveau terme n'ait été utilisé : de nouvelles entrées ont été créées et encodées de façon spontanée dans l'interface de catalogage du Centre, Pallas, lorsqu'aucun mot-clé préexistant ne semblait pertinent [11]. Hélas, faute de politique claire de gestion des métadonnées, ces nouvelles entrées n'ont pas été systématiquement contrôlées, traduites et standardisées. Or, la qualité des ajouts est très variable : il existe de nombreux problèmes de synonymie, d'ambiguïté, d'incohérences, de doublons et de non respect de la structure syntaxique requise. En outre, un sérieux problème de surindexation a été observé, certains descripteurs très pointus étant utilisés de façon tout à fait limitée pour indexer un document seulement (Béranger, 2017). Les principaux facteurs à l'origine de cette situation sont, d'après un rapport d'audit réalisé en 2015 (CegeSoma, 2015),

9. En ce qui concerne l'activité d'analyse documentaire et d'indexation des documents, précisons qu'elle ne semble pas systématique et uniformisée au CegeSoma. Ainsi, certains fonds sont parfois indexés à un niveau de granularité très fin – des mots-clés étant associés à une pièce en particulier –, tandis que d'autres fonds sont indexés uniquement à un niveau supérieur.

10. Sans parler des questionnements ocassionnés par l'intégration du CegeSoma aux Archives de l'Etat qui n'utilisent aucun langage documentaire de cet acabit.

11. Chaque catalogueur dispose en effet de la possibilité technique d'ajouter librement ses propres mots-clés.

le manque de formation, la négligence, l'absence de responsable du contrôle qualité, ainsi que les faiblesses et limites du logiciel d'encodage.

Concrètement, le thésaurus Pallas compte aujourd'hui [12] 120 841 mots-clés distincts. Seule une partie de ces descripteurs contient des noms de personnes [13]. Un sous-groupe composé uniquement de noms de personnes sera isolé lors des étapes de pré-traitement présentées au cours de la section suivante. Cependant, d'après une estimation grossière [14] nous pouvons déjà affirmer que ce jeu de données sera constitué au maximum de 28 000 noms de personnes [15], composés d'une seule chaîne de caractères pouvant être divisée jusqu'à cinq champs (en fonction du degré de précision et complétude de chaque descripteur) :

— nom
— prénom(s)
— année de naissance
— année de décès
— parfois une ou plusieurs informations complémentaires [16].

Ce sous-ensemble est inclu dans notre échantillon étant donné qu'il a été utilisé au cours de plus de deux décennies pour indexer les documents conservés par le CegeSoma et constitue ainsi une voie d'accès privilégiée aux collections. En outre, son caractère généraliste laisse penser qu'il pourrait rencontrer les intérêts d'un large public et que son intégration dans une instance Wikibase offrirait donc un intéressant *retour sur investissement* pour l'institution.

Deuxièmement, notre échantillon englobe un fichier contenant des informations sur 381 personnes impliquées dans des activités de résistance

12. L'export effectué à partir de la base de données Pallas et utilisé dans le cadre de cette étude de cas date du mois d'octobre 2019.

13. Ces derniers sont utilisés le plus souvent seuls (nom et prénom(s) accompagnés des années de naissance et de décès), mais il peut arriver qu'ils soient précisés à l'aide de termes juxtaposés, comme le montre ces exemples associés à Joseph Staline :
— staline, joseph (1879-1953) [vedette-matière seule]
— staline, joseph (1879-1953)–aspect politique [tête de vedette, subdivision thématique]
— staline, joseph (1879-1953)–espionnage–1898-1917 [tête de vedette, subdivision thématique, subdivision temporelle].

14. Basée sur l'utilisation d'un algorithme d'extraction de *noms potentiels* (voir Chardonnens *et al.*, 2018).

15. Ce nombre devra être revu à la baisse étant donné que cette première estimation inclue des faux positifs ayant été éronnément considérés comme des noms de personnes. C'est le cas par exemple de : *belgium, center christine de lalaing (1944)*. En outre, seuls les noms de personne associés à des archives et des photographies seront repris et utilisés dans cette étude de cas, étant donné que la migration des données du CegeSoma vers le logiciel de gestion des Archives a signifié la fin d'un traitement commun pour l'ensemble des collections : les livres et manuscrits sont désormais destinés à être pris en charge dans un système distinct.

16. Pseudonyme, épouse de, alias, titre de noblesse, etc.

en Belgique au cours de la Seconde Guerre mondiale et exécutées par l'occupant allemand entre 1940 et 1944 (CegeSoma, 2020b). Cette liste de personnes est issue d'un fonds documentaire [17] du CegeSoma contenant les lettres d'adieu écrites par des résistants [18] avant leur mise à mort. Cette liste, élaborée par des historiens du CegeSoma en vue d'une publication scientifique Maerten *et al.* (2011), contient un nombre variable d'informations sur chacun de ces suppliciés. Ce nombre peut aller jusqu'à un maximum de 27 éléments distincts (Maerten, 2013). Ces éléments englobent : nom ; prénom ; profession ; domicile ; région ; date de naissance ; état civil ; nombre d'enfants ; langue des lettres ; âge au moment de l'exécution ; moment de l'entrée en résistance ; fonction dans la résistance ; nom du groupement de résistance ; type de résistance ; date d'arrestation ; motif de condamnation ; date d'exécution ; lieu d'exécution ; pays d'exécution ; type d'exécution ; date des lettres ; nombre de lettres ; destinataires des lettres ; étendue de la restitution des lettres ; sources de la reproduction des lettres ; sources biographiques ; réalisation ou non d'une biographie et d'une analyse approfondie (Maerten, 2013).

Notre attention s'est portée sur ce jeu de données étant donné qu'il contient des informations beaucoup plus riches et détaillées que celles présentes par exemple dans les descripteurs du thésaurus du Centre. Bien qu'il soit rare que le personnel ait le temps de réaliser des notices aussi détaillées sur les personnes mentionnées dans des fonds d'archives et que ce jeu de données ne soit dès lors pas représentatif de la majorité des données préexistantes, la création de ce type de jeux de données plus détaillés apparaît toutefois comme une tendance à la hausse dans le contexte du CegeSoma. Dans cette optique, il semblait donc intéressant de pouvoir travailler avec des données empiriques afin de voir comment des données semi-structurées, parfois sujettes à ambiguïté [19], pouvaient être modélisées et implémentées dans le cadre d'une instance Wikibase. Finalement, précisons ici que les données réutilisées dans le cadre de notre échantillon se limitent aux données portant sur ces 381 personnes et n'incluent pas l'information relative aux lettres d'adieu qu'elles ont pu rédiger.

Enfin, un troisième jeu de données a été inclus à l'échantillon. Il s'agit de personnalités dont la biographie a été publiée sur la plateforme en ligne Belgium WWII, qui a été lancée par le CegeSoma en 2017, comme expliqué au cours du chapitre précédent. Ces notices synthétiques, publiées sur des pages

17. Désigné par la cote AA 2346.

18. Techniquement, certaines de ces personnes sont en réalité des otages (CegeSoma, 2020b) et non pas des membres actifs de la résistance.

19. C'est le cas par exemple des dates qui ont été encodées dans un format non normalisé ou des noms de lieux et de réseaux de résistance ayant été encodés en langage naturel.

individuelles [20] et disponibles en français, néerlandais et allemand, sont le fruit du travail collaboratif de nombreux scientifiques, à l'instar des autres contenus de la plateforme [21]. Bien qu'elles soient disponibles uniquement sous forme de texte publié en ligne et non sous la forme d'un fichier facilement réutilisable, ces données nous semblaient toutefois intéressantes dans le cadre de cette étude de cas. En effet, elles ont comme caractéristique commune de concerner des personnalités connues – ou relativement connues – de l'histoire de la Seconde Guerre mondiale en Belgique. Cette spéficité semblait particulièrement digne d'intérêt dans le cadre d'expérimentations liées à des processus de dédoublonnage de données ou de réconciliation d'entités nommées avec des ressources externes. En effet, cela permet de tester ces processus en ayant la probabilité d'avoir davantage de résultats – c'est-à-dire d'avoir des entités communes à plusieurs jeux de données – qu'en présence d'illustres inconnus. C'est pourquoi nous avons créé manuellement un jeu de données à partir des personnalités présentes sur le site Belgium WWII [22]. Ce fichier contient des informations sur 89 personnes liées à la Seconde Guerre mondiale en Belgique, il reprend leur nom ; prénom(s) ; année de naissance ; année de décès ; éventuels surnoms ou pseudonymes.

4.1.3 Respect de la vie privée

Il est clair que la publication de données portant principalement sur des personnes physiques soulève la question de la protection de la vie privée, notamment dans un contexte marqué par la récente entrée en vigueur du Règlement général sur la protection des données (RGPD) dans l'Union européenne. Dans le cadre de notre instance Wikibase, nous avons opté, sur conseil de la déléguée à la protection des données (DPO) des Archives de l'État, pour une publication de données restreintes à des personnes décédées [23], ce qui signifie que ces données ne sont dès lors pas soumises au RGPD [24].

Cependant, étant donné la présence dans les corpus concernés de nombreuses données ne comportant pas de dates de décès [25], nous avons pris la

20. Voir les pages Personnalités de la plateforme : https://www.belgiumwwii.be/belgique-en-guerre/personnalites.html.

21. La liste de l'ensemble des auteurs ayant contribué aux contenus du site web Belgium WWII est repris sur cette page : https://www.belgiumwwii.be/les-auteurs.html.

22. En date du mois de décembre 2019.

23. C'est également la stratégie adoptée par la base de données belge Odis (ODIS, 2020, p. 19), présentée au cours du premier chapitre et qui constitue l'une des sources d'inspiration de notre base de connaissance.

24. En effet, dans le Considérant n° 27, le Règlement précise que « this Regulation does not apply to the personal data of deceased persons » (GDPR.EU, 2020).

25. C'est le cas par exemple des listes nominatives associées à des interviews sur la vie quotidienne en Belgique durant l'Occupation.

décision d'élargir cette sélection à des personnes dont la date de naissance – remontant à 100 ans ou plus [26] – laisse présumer qu'elles sont aujourd'hui décédées. Pour couvrir le risque éventuel – bien que marginal et logiquement appelé à diminuer avec le temps – lié à la publication de ces données, la page d'accueil de la Wikibase fournit aux visiteurs la possibilité d'exprimer une plainte à ce sujet [27].

À l'avenir, il pourra être envisagé d'élargir cette sélection à des personnes vivantes faisant partie des exceptions prévues par le RGPD [28]. En effet, ce dernier prévoit des dérogations dans le cadre de la diffusion de données traitées à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques. Ces exceptions sont détaillées ainsi par le législateur belge, dans l'article 205 de la Loi du 30 juillet 2018 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel (Moniteur belge, 2018b) :

1. *La personne concernée a donné son consentement ; ou*
2. *les données ont été rendues publiques par la personne concernée elle-même ; ou*
3. *les données ont une relation étroite avec le caractère public ou historique de la personne concernée ; ou*
4. *les données ont une relation avec le caractère public ou historique de faits dans lesquelles la personne concernée a été impliquée.*

L'existence de ces exceptions pourrait en effet s'avérer utile au CegeSoma, notamment dans le cas de données concernant des figures de la collaboration en Belgique. Nous pouvons par exemple penser à des faits rendus publics par la personne concernée elle-même, comme dans le cas d'industriels ou collaborateurs désireux de se réhabiliter, tels De Coene de Courtrai, ou dans le cas de collaborateurs zélés fiers de leurs agissements comme Lode Claes [29].

26. Nous nous basons ici sur le délai traditionnellement utilisé au niveau fédéral en Belgique, découlant du délai de transfert stipulé dans la Loi sur les archives du 24 juin 1955 (Moniteur belge, 1955). Ce laps de temps est également utilisé dans le contexte archivistique flamand, par exemple par les archives de la ville d'Anvers (FelixArchief) Vanneste (2013). À titre indicatif, en Belgique, les délais de communication des registres d'état civil sont actuellement de 100 ans pour les actes de naissance, de 75 ans pour les actes de mariage et de 50 ans pour les actes de décès (il s'agit des délais en vigueur depuis 2019, suite aux modifications induites par l'article 166 de la Loi du 21 décembre 2018 (Moniteur belge, 2018a)).

27. Voir : https://adochs.arch.be/wiki/Main_Page#Vie_privée.

28. Cette réflexion pourrait notamment être nourrie des recommandations formulées par Dorien Styven, qui a examiné comment appliquer le RGPD dans le cadre des archives liées à l'Holocauste, et plus précisément dans le contexte belge de la Kazerne Dossin, voir Styven (2020).

29. Échanges par email avec Dirk Luyten, 06.08.2020.

Enfin, notons que si la possibilité de différer l'affichage d'une partie des données ou de limiter leur visibilité à certains types d'utilisateurs serait certainement utile dans ce type de contexte, une telle fonctionnalité n'est actuellement pas encore prise en charge par le logiciel Wikibase. En effet, Wikidata n'ayant pas été conçu dans l'optique de telles restrictions [30], le processus s'avérerait loin d'être trivial [31] étant donné les diverses fonctionnalités et extensions devant être prises en compte (comme l'historique, l'API, les données disponibles en JSON ou encore le service de requêtes SPARQL). Cela signifie que si l'institution tient à garder publiques les données pouvant être librement diffusées, elle devra alors trouver une solution alternative pour les données ne pouvant être librement diffusées (que cela soit via une base de données transitoire avant le décès d'une personne ou par l'utilisation d'une seconde Wikibase, avec tous les défis de synchronisation et de maintenance des données que cela implique).

4.2 Traitement

4.2.1 Pré-traitement

Le pré-traitement des données regroupe toutes les étapes de *nettoyage* et de préparation nécessaires avant de passer aux étapes ultérieures. En effet, si l'on souhaite par exemple lier toutes les occurrences du résistant belge René Robert en indiquant à l'aide d'un URI que toutes ces mentions font référence à un seul et même individu, il est nécessaire de pouvoir utiliser des informations connexes lors du processus de désambiguïsation. En effet, ce sont des *indices* comme son année de naissance (1894) ou son lieu de naissance qui nous permettront de vérifier si ces informations coïncident lorsqu'une occurence de *René Robert* apparaît. Or, ce travail de désambiguïsation est nettement facilité si l'information est encodée ou écrite de façon uniformisée. En effet, là où un être humain est par exemple capable de faire le lien entre *Liège*, *liege* et *la cité ardente*, une machine n'y verra que des chaînes de caractères bien distinctes. S'il est possible de favoriser cette opération en faisant appel à des algorithmes de type *fuzzy matching* permettant à la machine de rapprocher des termes selon leurs similitudes formelles, en revanche elle ne pourra pas établir seule que *la cité ardente* désigne la même ville que *Liège*, à moins de disposer de cette information. Il en va de même pour d'autres données comme les dates, qui sont parfois encodées de manière parfois bien

30. « But, at the end of the day MediaWiki was never designed with locking down in mind » [Message Telegram] Adam, 30.10.19.

31. « Just hiding edit links for example is probably rather trivial, while hiding certain things from some but not all users in the query service [...] is not trivial. », [Message Telegram], Lydia Pintscher, 29.10.2019.

différentes : 1943-09-08, 08-09-1943, ou encore 1943-9-8. Il est donc préférable de commencer par nettoyer et standardiser ce type de données (noms, prénoms, dates et lieux) afin de pouvoir plus aisément les comparer. Ce qui sera également bénéfique pour les étapes suivantes, visant à publier les informations sous forme d'URIs plutôt que de chaînes de caractères.

Sachant que l'encodage des données nominatives au CegeSoma est caractérisé par un contexte multilingue, un manque de standardisation et des pratiques globalement peu harmonisées, d'importants efforts de pré-traitement sont donc requis. Ce travail est d'autant plus conséquent qu'il concerne différents fichiers et que la méthode diffère selon le type de fichiers de départ. En effet, comme souligné dans la section précédente, un fichier tabulaire disposant de champs distincts ne requiert par exemple pas la même étape de *parsing* qu'un fichier de données contenant plusieurs informations de nature distincte dans la même chaîne de caractères.

Afin de limiter les redondances susceptibles d'alourdir la lecture de ce texte, une vue d'ensemble des différentes étapes effectuées est présentée ici, sans aborder le détail du traitement de chaque fichier en particulier. Cependant, ce travail de pré-traitement ne doit pas être sous-estimé. En effet, le principe de Pareto se manifeste également dans ce contexte et, comme Dasu et Johnson (2003) l'ont par exemple montré, il apparaît que les *data scientists* dédient environ 80% de leurs efforts à la collecte, au nettoyage et à la réorganisation des données, et que 20% seulement sont dévolus aux analyses en tant que telles. Il semble donc crucial de ne pas laisser dans l'ombre ces opérations de pré-traitement, en exposant les difficultés rencontrées, aussi triviales soient-elles. Des détails sont donnés sur les quatre étapes principales ayant été utilisées pour traiter les noms, prénoms, lieux et dates.

La première étape a pour objectif d'homogénéiser les données en gommant de petites différentes formelles. Elle consiste à supprimer les espaces inutiles (par exemple en début et en fin de chaîne de caractères, ou alors deux espaces consécutives), à passer provisoirement tous les caractères en minuscules et à remplacer les caractères avec accents et cédille par des caractères sans accents. Cette opération pouvant être facilement automatisée et réalisée en quelques secondes sur des milliers de lignes, il ne semble donc pas nécessaire de la détailler davantage.

La seconde étape, dédiée au *parsing*, vise à analyser un flux de caractères pour en extraire certains éléments. Dans notre cas, le but est d'extraire des éléments de même nature comme des noms ou des dates, pour les isoler dans une cellule distincte. Réalisée à l'aide du logiciel OpenRefine, cette opération a été réalisée en tenant compte de la présence de motifs récurrents, repérables à la présence de certains caractères-clés comme par exemple des signes de ponctuation. Par exemple, la plupart des mots-clés

Pallas sont construits selon cette structure *nom, prénom(s) (année-année)*, c'est-à-dire une première chaîne de caractères correspondant au nom de famille, suivie d'une virgule, elle-même suivie du ou des prénom(s), suivis d'une parenthèse ouvrante, suivie de quatre chiffres correspondant à l'année de naissance, suivi d'un tiret et de quatre nouveaux chiffres correspondant à l'année de décès, et d'une parenthèse fermante. Le repérage de ces motifs permet de traiter la plupart des cas et d'ainsi passer d'une seule chaîne de caractères à un ensemble de colonnes regroupant des données de même nature : nom, prénom(s), année de naissance, année de décès.

Cependant, il faut savoir que cette étape n'est pas la plus chronophage : si toutes les données suivaient cette structure, ce travail serait rapidement achevé et il n'y aurait rien de plus à en dire. C'est le traitement de tous les cas particuliers qui est le plus gourmand en temps et nécessite certaines opérations sur mesure. En effet, tous les noms ne suivent pas le même schéma et il est apparu que des informations de nature tout à fait variée ont été ajoutées au schéma classique [Nom, prénom, (date-date)]. Voici quelques exemples de ces cas particuliers rencontrés au sein des mots-clés Pallas :

Alias *faust, camille laurent celestin (alias mauclair, camille ; 1872-1945)*

Nom de naissance *schoeffler, anna (b. hauptmann ; 1898-144)*

Surnom *genotte, fernand (dit nandy ; 1923-1961)*

Épouse de *mette, jeanne (epouse mendes, catulle ; 1867-1965)*

Titres honorifiques et de noblesse *bazy, louis (marquis de mun ; 1883-1960)*

Prénom seul et titre *oskar (prince de prusse ; 1888-1958)*

Date inconnue *castan, julien (b.-1993)*

Date incomplète *enoch, paul (19..-1970)*

Ces exemples montrent la façon dont la parenthèse ouvrante débouche parfois sur une autre information que la date de naissance, comme par exemple un alias. Il a ainsi été nécessaire de commencer par traiter ces cas particuliers, afin d'améliorer les résultats du *parsing* en isolant les informations additionnelles. Dans la mesure du possible, ces informations ont été placées dans une colonne dédiée selon leur nature, afin de pouvoir être réutilisées et valorisées ultérieurement.

Par ailleurs, le *parsing* peut également être utilisé afin de séparer les éventuels deuxième ou troisième prénoms dans des colonnes dédiées. Le fait d'isoler l'information selon sa nature permet d'affiner la recherche de *doublons* (deux chaînes de caractères faisant référence à la même personne) en comparant ce qui est comparable : un nom de famille accompagné d'un

prénom seul (Dupont Jeanne) diffère fortement d'un nom de famille accompagné d'un prénom suivi d'initiales ou d'un second prénom (Dupont Jeanne Marie) et résultera en un score de *matching* probablement peu élevé, alors qu'il s'agit peut-être du même individu, la première occurrence étant simplement moins complète que la seconde. Ce second prénom peut toutefois constituer un élément utile lors du processus de désambiguïsation et ne doit dès lors pas être négligé, simplement, le processus gagnera en clarté si ce second prénom est déplacé dans une colonne dédiée.

Finalement, dans d'autres cas encore [32], il arrive que des méthodes *artisanales* aient été utilisées afin de signifier des graphies alternatives. C'est le cas par exemple de certains noms de famille, comme *Col(l)inet*, dont l'encodage indique une incertitude entre *Colinet* et *Collinet*, mais cela survient également pour des prénoms variant selon les langues, comme *Julianus (Juliaan ou Julien)* ; *François (Frans)* ; *Joannes Baptista (Jan Baptist)*. Évidemment, moins la méthode est standardisée, plus cela exige de procéder manuellement pour gérer ces occurrences, comme le montre par exemple ce contenu, *Maurice ? Moïse ?*, qui ne suit pas la structure – implicite – du reste de la colonne prénoms. Ces cas particuliers ont été traités en créant une nouvelle colonne dédiée aux graphies alternatives.

La troisième étape concerne les dates, notre objectif étant de les harmoniser en utilisant la norme internationale ISO 8601 [33] comme référence. Il s'agit donc de convertir toutes les dates vers cette notation, quel que soit leur niveau de granularité. À nouveau, des outils permettent d'automatiser facilement cette conversion en précisant le format de sortie souhaité. La complexité survient lorsque les données initiales contiennent des problèmes de qualité. Ainsi, l'un des jeux de données [34] contenait des dates encodées de deux manières différentes : 04/21/1920 et 23/10/1891 ; la première est construite selon le format américain *mm/dd/yyyy*, dit *middle-endlan*, tandis que la seconde est écrite selon le format en vigueur en Europe : *jj/mm/aaaa*. Les jours et les mois sont donc intervertis. S'il est possible de déduire où sont placés les jours et où sont placés les mois lorsqu'il s'agit de dates au-delà du 12 du mois (comme par exemple 24/12/1944), cela s'avère plus ardu lorsqu'on est en présence d'une date telle que le 04/02/1920. En effet, cela pourrait tout aussi bien être le 2 avril 1920 que le 4 février 1920, en fonction de la notation utilisée. Cette absence de standardisation au sein d'un même fichier a pu être vérifiée et corrigée manuellement dans ce cas pré-

32. Par exemple dans le cadre du fichier nominatif associé au fond AA2346 (qui concerne les lettres d'adieu des résistants de Belgique exécutés en 1940-1944).

33. Voir https://www.iso.org/fr/iso-8601-date-and-time-format.html.

34. du fichier nominatif associé au fond AA2346 (qui concerne les lettres d'adieu des résistants de Belgique exécutés en 1940-1944).

sent [35], cependant il en irait autrement avec des fichiers ne contenant pas de tels signes distinctifs et plusieurs milliers de lignes. Par ailleurs des notations incomplètes comme par exemple 6/3/1900 compliquent l'opération de conversion automatisées, nécessitant au préalable de compléter la date avec les zéros manquants.

Enfin vient la quatrième et dernière étape : les noms de lieux. Permettant par exemple de préciser le lieu de naissance d'une personne, ils doivent également faire l'objet d'un pré-traitement, afin de pouvoir être comparés de façon optimale au cours des prochaines étapes – favorisant ainsi le processus de désambiguïsation des noms de personnes – et plus facilement réutilisés par la suite. Ils recèlent une complexité particulière, se manifestant à différents niveaux. Ainsi, une chaîne de caractères comme *Alost*, mentionnée sans autre information complémentaire, peut désigner des entités de lieux tout à fait distinctes. Cela s'explique par l'existence de plusieurs *Alost* belges à travers l'espace, mais aussi à travers le temps. À travers l'espace, car il arrive que des localités distinctes portent le même nom. Ainsi, Alost désigne à la fois une ville dans la province de Flandre-Orientale et une section de la ville de Saint-Trond, dans la province du Limbourg. Mais on parle également de changements à travers le temps, car le sort des lieux évolue au gré des événements historiques et politiques. Ainsi, si l'on se concentre sur le XX^{e} siècle, période au cœur des collections du CegeSoma, il faut bien entendu prendre en compte l'important processus de fusion des communes belges, débuté en 1977 et encore en cours aujourd'hui [36]. Ces fusions se traduisent notamment par des changements de noms de lieux. Ainsi, la commune d'Alost avant fusion est devenue l'une des sections de la commune actuelle de Alost, qui englobe également Gijzegem, Hofstade, Baardegem, Herdersem, Meldert, Moorsel, Erembodegem, Nieuwerkerken. Pour éviter toute ambiguïté, il est donc nécessaire de pouvoir clairement signifier si l'on fait référence à Alost l'ancienne commune – actuelle section de la commune unifiée Alost –, ou à la commune actuelle, après fusion. Enfin, il faut également prendre en compte la dimension du multilinguisme : Alost est la forme en français de Aalst (en néerlandais). Sans compter que des graphies alternatives peuvent exister, issues par exemple de dialectes flamands ou wallons.

Cette complexité signifie qu'une simple chaîne de caractères ne permet pas d'identifier de façon claire et unique un lieu. Il est dès lors nécessaire de faire appel à un référentiel préexistant muni d'identifiants uniques et d'informations complémentaires comme des dates et des coordonnées géographiques, afin de pouvoir désigner un lieu en levant toute ambiguïté. Dès

35. Des détails dans le fichier d'origine – alignement à gauche ou alignement à droite de la cellule Excel – ont permis de distinguer les deux types de notation et de les adapter en suivant la norme ISO 8601.

36. La Flandre est ainsi passée de 308 à 300 communes le 1er janvier 2019 (Statbel, 2019).

lors qu'un tel référentiel est utilisé pour l'encodage de nouvelles données, il devient aisé de regrouper des entités similaires : Alost, la section de commune de Saint-Trond, dans la province de Limbourg, étant alors désignée par le même identifiant unique que la forme néerlandaise *Alast* ; cette entité lieu étant désormais clairement distincte des autres Alost, qu'il s'agisse de la ville située en Flandre orientale ou de l'une des sections de cette commune, possédant la même dénomination.

Les bases de connaissance Wikidata [37] ou GeoNames [38] semblaient toutes indiquées pour mener à bien cette tâche, en faisant appel aux URIs des lieux concernés [39]. Cependant, l'observation des données, la prise en considération des particularités liées à la fusion des communes, ainsi que des échanges avec le gestionnaire des bases de données des Archives de l'État nous ont poussée à privilégier le référentiel qu'utilise l'institution en interne, ce dernier s'avérant plus exhaustif et adapté à nos besoins. Cette utilisation semble d'autant plus stratégique dans un contexte où le CegeSoma s'emploie actuellement à adapter ses pratiques à celles des Archives de l'État.

Concrètement, chaque entité du référentiel est accompagnée de son code INS [40] – ce qui n'était pas le cas de la majorité des entités Wikidata concernées –, mais également de coordonnées géographiques – ce qui n'est pas le cas des fichiers mis à disposition par l'office belge de statistiques. De plus, la co-existence des anciennes communes et communes unifiées est plus systématique ; le fichier est régulièrement mis à jour afin de tenir compte des modifications issues de nouvelles fusions de communes et enfin, il permet de faire face à un large cas de figure, étant donné qu'il comporte également des noms de hameaux et lieux-dits. Ces constats en faveur du référentiel des Archives de l'État n'enlèvent toutefois pas l'intérêt d'une base de connaissance telle que Wikidata, qui dispose par exemple de beaucoup plus de traductions et graphies alternatives, ainsi que d'informations complémentaires, comme la taille de la population à une année donnée. En marge des étapes de prétraitement des données du CegeSoma, des efforts ont ainsi été déployés afin de compléter à la fois les entités Wikidata et le référentiel des Archives de l'État, les fruits de ce labeur étant évidemment destinés à être ensuite réutilisés dans la future Wikibase dédiée aux entités Personne du CegeSoma. Le détail de ces opérations est repris dans l'Annexe 3 [41].

37. https://www.wikidata.org/.
38. https://www.geonames.org/.
39. Par exemple : Q2569898|Céroux-Mousty (https://www.wikidata.org/entity/Q2569898 ; 2800584|Céroux-Mousty (https://www.geonames.org/2800584/ceroux-mousty.html).
40. Ce code numérique composé de cinq caractères est attribué à chaque entité administrative belge par Statbel, l'office belge de statistiques Statbel (2018).
41. Page 334.

Ces étapes préliminaires sur les noms de lieux terminées, il fut temps de traiter les noms de lieux présents dans les différents jeux de données à disposition. Divers types de problèmes se sont présentés au cours de nos efforts de réconciliation. Par exemple, *Ougrée* est une ancienne commune qui a été coupée en deux ; faut-il la lier à la *partie de Ougre* associée à la commune de Seraing ou la lier à celle de Liège ? *Fouron-le-Comte* s'avère être la forme en français de 's Gravenvoeren, mais encore faut-il que l'information soit disponible pour effectuer ce lien d'équivalence. *Ninane* ne *matche* avec aucune entité du référentiel. Il faut effectuer quelques recherches supplémentaires pour comprendre qu'il s'agit d'un village faisant partie de Chaudfontaine. Pas de résultat pour *Altine* ; s'agit-il d'Haltinne, qui se dit Altene en wallon ? Si l'on n'est pas fin connaisseur de la géographie belge, une véritable enquête doit être menée pour découvrir que *Sint-Pieters (bij Brugge)* correspond à Sint-Pieters-op-den-Dijk, un quartier de Bruges, qui fut une commune indépendante jusqu'en 1899 et fait aujourd'hui partie de Bruges... mais ça ne nous dit pas à quelle commune ou section de commune de Bruges il faut l'associer ? Quant à *St-Antonius-Brecht*, ce village faisait partie de la commune de Brecht avant d'être ajouté à Zoersel le 1er janvier 1977. Par ailleurs, comment traiter des cas comme *Moresnet* [42], un territoire communal neutre avant 1918, avant de devenir une commune belge en 1919 ? Sans parler des cas où il serait tellement utile de pouvoir se plonger dans la tête de la personne encodant les données, s'agit-il d'*Hofstade*, à Alost ou dans le Brabant flamand ? Ces deux sections de commune étant situées à 50 kilomètres l'une de l'autre, ce n'est pas un détail. Une investigation plus poussée sur la personne associée à ce lieu de naissance permet certes de déduire qu'il s'agit de Hofstade à Zemst, mais il s'agit d'une tâche extrêmement chronophage, difficilement envisageable pour des centaines voire des milliers de cas ambigus.

Il n'existe pas de solution unique pour gérer ces différents cas de figure et certains, trop complexes pour être résolus rapidement, ont ainsi dû être temporairement laissés de côté. Les exemples énumérés dans le paragraphe précédent ont donc le mérite de mettre en lumière à la fois les limites de l'automatisation et la nécessité de mettre en place des protocoles guidant le personnel dans l'utilisation d'URIs à la place de chaînes de caractères qui, bien souvent, sont synonymes d'ambiguïté.

Au terme de ces quatre étapes de pré-traitement, les données sont maintenant prêtes pour les étapes ultérieures, à savoir les opérations d'*entity linking*, de réconciliation et d'enrichissement, décrites et illustrées à l'aide d'exemples au cours des prochaines sections.

42. Merci à la volontaire du CegeSoma qui a fait remonter ce cas particulier jusqu'à nous.

4.2.2 Entity linking

Le travail d'intégration de jeux de données disparates au sein d'une seule infrastructure de gestion et de publication des données requiert de pouvoir d'abord repérer et fusionner les *doublons*, afin d'assurer la meilleure qualité de données possible. Cela implique de commencer par associer les valeurs faisant référence à une même personne. Ce processus fait partie du *data matching*[43] : « the task of identifying, matching, and merging records that correspond to the same entities from several databases » (Christen, 2012, p. ix).

La difficulté de cette tâche réside dans l'absence d'identifiants communs entre ces différentes bases de données. Il ne s'agit donc pas de se reposer uniquement sur les chaînes de caractères des noms ou prénoms, mais également de faire appel à des éléments utiles à l'identification d'une entité (Christen, 2012). Dans notre cas, il s'agit par exemple de dates de naissance et de décès, et, lorsqu'ils sont présents, de lieux de naissance ou de décès. Cependant, comme nous l'avons vu, ces données sont sujettes à des problèmes de qualité, et il n'est pas rare de constater des fautes de frappe, des variations d'orthographe, des informations incomplètes ou ayant évolué dans le temps.

Pour effectuer ces opérations, nous avons opté pour une bibliothèque Python d'*entity linking* : the Python Record Linkage Toolkit[44], qui permet de lier les valeurs d'une ou de plusieurs sources de données. Utilisant Pandas et Numpy, cette bibliothèque englobe tous les outils et algorithmes nécessaires pour dédupliquer et lier des données en mesurant la similarité entre plusieurs variables, qu'il s'agisse de nombres, de chaînes de caractères ou de dates. En outre, elle est accompagnée d'une documentation illustrée de nombreux exemples[45], et elle a également fait l'objet de tutoriels approfondis[46] permettant de faciliter sa prise en main. Quelques tests préliminaires ayant confirmé son adéquation à nos besoins, ce module d'*entity linking* a donc été utilisé pour l'entièreté de ces opérations.

Ce processus d'*entity linking* peut être décomposé en trois principales étapes[47] :

43. Également connu sous d'autres noms, comme : *data matching*, *entity linkage*, *data linkage*, *entity resolution*, *object identification*, *field matching*, ou encore *deduplication* lorsqu'il s'agit de valeurs issues d'une même source de données.

44. Voir le dépôt GitHub associé au projet https://github.com/J535D165/recordlinkage.

45. http://recordlinkage.readthedocs.org/en/latest/.

46. Voir par exemple cet ensemble de tutoriels du laboratoire de recherche Netlab : https://uwaterloo.ca/networks-lab/blog/post/pre-processing-recordlinkage.

47. En réalité, la documentation inclut également une étape préliminaire de pré-traitement afin de pallier certains problèmes de qualité des données (nous ne l'incluons cependant pas ici, cette première étape ayant fait l'objet de la sous-section précédente), de même qu'une étape d'évaluation (réalisée manuellement dans notre cas).

1. Indexation
2. Comparaison
3. Classification

La première étape, l'indexation, également parfois appelée *blocking* (Christen, 2012), consiste à créer les paires de données faisant potentiellement référence à la même entité, afin de pouvoir les comparer. Ces paires constituent des *candidate links*. En toute logique, plus le nombre de lignes à traiter augmente, plus cette étape est conséquente. Des algorithmes de *blocking* ont donc vu le jour, afin de réduire la vitesse de calcul et de limiter les comparaisons aux cas les plus plausibles.

En effet, comme l'explique Becker, « in most data integration problems, the majority of possible links will be non-matches, and so we want to use our knowledge of the data we're integrating to eliminate bad links from the outset » (Becker, 2017). La méthode d'indexation basée sur le *blocking* vise donc à générer uniquement des paires de candidats lorsque les données coïncident au niveau d'une ou de plusieurs variables (par exemple, il est possible de ne générer des *candidates links* que lorsque le nom de famille de la personne commence par la même première lettre). Il faut noter que cette méthode nécessite une bonne connaissance des données afin d'éviter d'exclure par erreur de véritables doublons : « the ideal indexing strategy will discard a large number of mismatching records but discard very few matching records » (Becker, 2017).

Pour accomplir cette étape de façon optimale, ainsi que les deux suivantes, nous avons opté pour l'élaboration d'un *gold standard corpus*, qui identifie les véritables paires concordantes contenues parmi toutes les paires potentielles. Si l'élaboration manuelle d'un tel corpus est loin d'être optimale [48], elle paraît toutefois utile dans le cadre de cette étude de cas basée sur des données empiriques possédant des caractéristiques propres et n'ayant jamais fait l'objet d'opérations d'*entity linking* jusque-là. Par ailleurs, bien que le *gold standard corpus* soit généralement utilisé au terme du processus, lors d'une étape d'évaluation des résultats [49], nous l'avons utilisé à chacune des étapes, afin d'affiner au fur et à mesure les méthodes utilisées [50] et de vérifier que les *vrais positifs* n'étaient pas exclus de manière involontaire.

48. Ainsi, outre le fait qu'il caractérise ce processus de classification manuelle comme long, fastidieux et sujet aux erreurs (Christen, 2012, p. 34), Christen met en garde contre son utilisation comme *gold standard* qui peut être *dangereuse* (Christen, 2012, p. 35).

49. Comme l'explique Christen : « to evaluate the completeness and accuracy of a data matching project, some form of ground-truth data, also known as gold standard, are required. Such ground-truth data must contain the true match status of all known matches (the true non-matches can be inferred from them) » (Christen, 2012, p. 34).

50. Comme le soulignent Dusetzina *et al.*, « both deterministic procedures and probabilistic procedures should be considered iterative. After completing the initial linkage, a random

Ainsi, en ce qui concerne l'étape d'indexation, cela nous a par exemple permis de tester la méthode de *blocking* à l'aide de différentes variables et de constater qu'une indexation restreinte aux paires contenant un nom de famille identique [51] nous permettait d'obtenir l'intégralité des vrais positifs repérés au sein de notre *gold standard* [52], tout en réduisant significativement la quantité de paires à comparer.

La seconde étape, de comparaison, vise à définir des éléments précis devant être comparés afin de pouvoir retrouver quels *candidate links* correspondent à une même entité (il s'agira par exemple de comparer le degré de similitude entre deux années de naissance). Pour y parvenir, différentes méthodes de comparaison [53] peuvent être utilisées conjointement, qu'elles soient basées sur des similarités strictes ou encore sur la distance d'édition entre deux chaînes de caractères. Cette étape est directement conditionnée par la richesse des jeux de données destinés à être comparés : plus les noms de personnes sont accompagnés de données discriminantes, plus le processus de désambiguïsation pourra être affiné et plus les résultats pourront être considérés avec confiance. Il s'agit donc d'observer quels sont les attributs communs aux deux fichiers et de déterminer quelle sera la méthode la plus adaptée pour procéder à la comparaison de ces attributs.

En ce qui concerne nos données, nous avons relevé l'intérêt de mener différentes expérimentations afin d'affiner le choix des méthodes de comparaison utilisées [54]. Cela nous a également permis de tester des méthodes plus expérimentales comme cet algorithme phonétique [55] permettant de conver-

sample of match decisions should be reviewed to ensure that the algorithm is performing as intended. If the review process reveals opportunities for improvement, then the algorithm should be adjusted to account for the identified weaknesses » (Dusetzina *et al.*, 2014).

51. Cette variable est appropriée ici étant donné que les données possèdent une qualité suffisante pour se reposer sur les noms de famille, si ce n'était pas le cas, il pourrait alors être intéressant de recourir à un algorithme offrant davantage de flexibilité comme *SortedNeighbourhood* (https://recordlinkage.readthedocs.io/en/latest/ref-index.html#recordlinkage.index.SortedNeighbourhood).

52. Le détail du code ayant permis de mener à bien ces opérations est repris dans l'Annexe 4, page 345.

53. Pour une vue exhaustive, voir : https://recordlinkage.readthedocs.io/en/latest/ref-compare.html#module-recordlinkage.compare.

54. Cela nous a permis par exemple de repérer qu'une méthode comme *Compare.Numeric* (https://recordlinkage.readthedocs.io/en/latest/ref-compare.html#recordlinkage.compare.Numeric) permet par exemple une comparaison beaucoup plus fine que ne le permet *Compare.Exact* (https://recordlinkage.readthedocs.io/en/latest/ref-compare.html#recordlinkage.compare.Exact). Prenons par exemple une personne répondant au nom de Charles Rahier, née en 1909 selon la page Personnalités du site Belgium WWII (https://www.belgiumwwii.be/belgique-en-guerre/personnalites/charles-rahier.html) et née en 1910 selon le thésaurus Pallas du CegeSoma : le score de comparaison de l'année de naissance associée à cette entité est de 0.5 avec la méthode *Compare.Numeric*, tandis qu'il est de 0.0 avec la méthode *Compare.Exact*.

55. Voir : https://recordlinkage.readthedocs.io/en/latest/ref-preprocessing.html#phonetic-encoding.

```
Entrée [13]: df11.head()
```

Out[13]:

	B_name	B_prenom	B_nom	birth	death	B_wikidata_id	cleaned_prenom	cleaned_nom	phoneticPrenom	phoneticNom
0	Achille Van Acker	Achille	Van Acker	1898.0	1975.0	Q14997	achille	van acker	ACAL	VANACAR
1	Adrien Emile Van Coppenolle	Adrien Emile	Van Coppenolle	1893.0	1975.0	NaN	adrien emile	van coppenolle	ADRANANAL	VANCAPANAL
2	Albert Lilar	Albert	Lilar	1900.0	1976.0	Q466832	albert	lilar	ALBAD	LALAR
3	Albert Servaes	Albert	Servaes	1883.0	1966.0	Q2197592	albert	servaes	ALBAD	SARV
4	Alexander von Falkenhausen	Alexander	von Falkenhausen	1878.0	1966.0	Q62521	alexander	von falkenhausen	ALAXANDAR	VANFALCANASAN

FIGURE 4.2 – Aperçu d'un jeu de données, complété à l'aide de codes phonétiques issus des prénoms et noms, dans le cadre du processus d'*entity linking*.

```
Entrée [40]: matches3.head()
```

Out[40]:

		N_exact	N_jarowink	P_exact	P_jarowink	P_Phon_exact	N_Phon_exact	date1_exact	date1_numeric	date2_exact	date2_numeric	score
belgium	adieux											
60	28	0	0.0	0	0.0	0	0	1	1.0	1	1.0	4.0
66	66	0	1.0	0	0.0	0	0	0	0.0	1	1.0	3.0
	65	0	1.0	0	0.0	0	0	0	0.0	1	1.0	3.0
50	170	0	1.0	0	0.0	0	0	0	0.0	1	1.0	3.0
61	79	0	1.0	0	0.0	0	0	1	1.0	0	0.0	3.0

FIGURE 4.3 – Aperçu du système de scores permettant de comparer des *candidates links* sur la base de différentes variables.

tir des noms en codes phonétiques sur la base de leur prononciation [56], comme l'illustre la figure 4.2.

Comme souligné, cette étape varie énormément en fonction des attributs susceptibles d'être comparés. Dans certains cas, seuls les noms, prénoms et années de naissance et de décès étaient disponibles. Notre démarche a consisté à comparer ces valeurs à l'aide de différents algorithmes : ainsi, comme le montre la figure 4.3, le score obtenu tient compte, d'une part, de la similarité parfaite [57] entre deux prénoms, mais également entre les deux codes phonétiques associés à ces prénoms, et, d'autre part, de la similarité partielle [58] entre ces prénoms, basée sur la mesure de distance de Jaro-Winkler [59]. Si un processus itératif tenant compte du *gold standard corpus* permet de perfectionner l'utilisation de ces méthodes de comparaison afin d'arriver à des résultats tout à fait convaincants [60], il faut cependant garder à l'esprit qu'il suffit de changer de corpus et d'être confronté à d'autres types de données – comme par exemple un nom de lieu ou un nom de réseau de résistance – pour que le modèle doive être à nouveau ajusté.

56. Nous devons préciser que ce processus n'est cependant pas optimal dans la mesure où l'algorithme du *Record Linkage Toolkit* que nous avons utilisé, *NYSIIS* (https://jellyfish.readthedocs.io/en/latest/phonetic.html#nysiis), se base sur la prononciation anglaise des noms.

57. Compare.Exact, voir : https://recordlinkage.readthedocs.io/en/latest/ref-compare.html#recordlinkage.compare.Exact.

58. Compare.String, voir : https://recordlinkage.readthedocs.io/en/latest/ref-compare.html#recordlinkage.compare.String.

59. C'est la mesure de distance qui a fourni les résultats les plus convaincants lors de tests préliminaires.

60. En témoigne un F-score – mesure destinée à tester la précision et le rappel – de 0.967.

Une fois les *candidate links* ayant pu être comparés sur la base d'une série d'attributs communs, il s'agit d'établir un seuil permettant de distinguer les paires concordantes *(possible matches)* des paires non concordantes *(non-matches)* (Anderson, Jillian, 2018) ; il arrive également que soient incluses dans la classification les concordances potentielles *(potential matches)* (Christen, 2012). Par ailleurs, si une méthode basée sur un seuil nous semblait suffisante dans la présente situation, il faut noter que les méthodes de classification incluent également des modèles d'apprentissage (*machine learning)* tant supervisés que non supervisés [61].

Concrètement, cette étape consiste à calculer les scores obtenus au cours de l'étape de comparaison et à établir un score minimal permettant de générer un nouveau fichier à partir des paires candidates, en excluant les paires ne possédant pas suffisamment de points communs. Cet ensemble inclut à la fois les concordances *parfaites* et les concordances potentielles, qui nécessiteront une vérification manuelle. Dans les cas où le temps et les ressources manqueraient pour effectuer de telles vérifications, ce score pourrait alors être utilisé afin de caractériser le taux de confiance ayant conduit à associer deux entités distinctes, à l'instar de ce qui a été mis en place par the National Archives UK dans le cadre du projet Traces through Time (Ranade, 2016a).

Enfin, notons que notre travail portant sur un échantillon limité, nous avons choisi par commodité de réaliser la suite des opérations *manuellement* [62]. Cependant, une étape comme la fusion des paires concordantes gagnerait toutefois à être automatisée dans le cadre d'opérations réalisées à plus large échelle, afin de limiter le nombre d'opérations de manipulation de données, qui accroît le risque d'erreurs.

4.2.3 Réconciliation

Ce processus vise à associer un URI correct (par exemple issu d'une base de connaissance comme Wikidata) à chaque entité nommée, de manière à les désambiguïser [63]. Étant donné la quantité parfois très limitée de données d'identification attachées à chaque entité, ce processus de réconciliation est à la fois complexe et utile.

Pour commencer, nous devons préciser que si de multiples ressources externes, tant généralistes que spécialisées, ont été envisagées dans un premier

61. Voir : https://recordlinkage.readthedocs.io/en/latest/ref-classifiers.html.

62. Plus concrètement, à l'aide du logiciel OpenRefine et de la fonction *join* permettant d'inclure des colonnes issues d'autres projets.

63. Comme le recommandent Casalini *et al.*, idéalement cette étape devrait déjà être effectuée au moment de la création des données : « When creating new authority records manually, data creators are encouraged to consult large data aggregations and use their identifiers in their local records to actively disambiguate their identity. » (Casalini *et al.*, 2018, p. 18), ce n'était cependant pas le cas au CegeSoma.

temps [64], il s'est avéré, d'une part, que toutes n'offraient pas une voie d'accès privilégiée à leurs données [65], et, d'autre part, qu'au vu des ressources conséquentes que requièrent de tels alignements, il semblait plus stratégique de concentrer dans un premier temps nos efforts sur Wikidata. En effet, cette base de connaissance se démarque par son rôle de *linking hub* : cela signifie que se concentrer sur elle pour l'étape de réconciliation permet de récupérer simultanément l'URI Wikidata et d'autres identifiants externes possédant une propriété Wikidata dédiée [66].

Concrètement, ce sont les fonctionnalités de réconciliation proposées par le logiciel libre OpenRefine qui ont été utilisées pour procéder à cet alignement. Étant donné qu'il existe déjà des tutoriels détaillés à ce sujet [67], nous ne présentons pas ici le détail des opérations [68]. Cette étape nous a permis de mener à bien ce travail pour une partie de notre échantillon, de manière à ce que les identifiants Wikidata puissent être documentés dans notre Wikibase et éventuellement être utilisés comme pivots pour récupérer d'autres identifiants externes ou des données complémentaires. Nous avons également posé les jalons pour le jeu de données le plus volumineux, issu du thésaurus Pallas, de manière à ce que les étapes de vérification manuelle puissent être réalisées par une équipe dédiée au cours des prochains mois. Comme le montre le tableau 4.1, qui propose une vue récapitulative de ces premières étapes de réconciliation, sur les près de 30 000 potentiels noms de personnes identifiés au sein du thésaurus Pallas – qui contient au départ un total de près de 120 000 descripteurs –, plus d'un tiers d'entre eux ont été associés à des candidats Wikidata, avec un score de réconciliation supérieur ou égal à 50%.

Le tableau précise le nombre de potentiels noms de personnes accompagnés de dates, étant donné que les années de naissance et de décès ont été ajoutées comme source d'information additionnelle lors du processus de réconciliation en vue d'améliorer la qualité et la finesse des résultats. Pour al-

64. Comme Wikidata, Virtual International Authority File, Social Network Archival Context, International Standard Name Identifier, Centro di Documentazione Ebraica Contemporanea, NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies, European Portal Archive, Institut für Zeitgeschichte, the Belgian War Dead Register, WWI the name lijst ou encore Mémoire des hommes.

65. Par exemple par le biais d'une API de réconciliation.

66. Comme par exemple P213|identifiant ISNI (https://www.wikidata.org/wiki/Property:P213).

67. Voir par exemple la documentation proposée par OpenRefine (https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation) ou encore ce tutoriel proposé par Mathieu Saby (https://msaby.gitlab.io/formation-openrefine-Lyon20191122/enrichir-ses-données-et-aligner-sur-des-sources-extérieures.html#alignement-de-données-avec-des-réfrentiels-externes).

68. Par ailleurs nous renvoyons également les lecteurs vers l'Annexe 1, page 323, qui traite du processus similaire de réconciliation effectué dans le cadre de l'analyse des requêtes d'utilisateurs du catalogue en ligne Pallas et en souligne certaines limites.

Taille initiale du fichier (lignes)	118 682
Potentiels noms de personnes	28 457
Potentiels noms de personnes avec dates	20 466
Score de réconciliation Wikidata $\geq 50\%$	9 694

TABLE 4.1 – Résultats préliminaires de la réconciliation des noms de personnes issus du *thésaurus Pallas* avec Wikidata

ler plus loin, il pourrait être envisagé d'utiliser d'autres types d'informations additionnelles, comme par exemple des entités nommées présentes dans le titre du document ou dans la légende de la photo à laquelle a été associé un descripteur.

En ce qui concerne les cas où aucune réconciliation convaincante n'a pu être effectuée, trois cas de figure sont à envisager.

Premièrement, il peut s'agir de noms de personnes qui sont trop communs pour être aisément désambiguïsés [69] et sont associés à trop de noms candidats : il est alors nécessaire d'effectuer une vérification manuelle en consultant d'autres sources d'informations pour trancher. Deuxièmement, les noms de personnes peuvent appartenir à du contenu de niche qui n'a pas encore été décrit par des membres de la communauté Wikidata : cela pourrait alors valoir la peine d'explorer d'autres ressources issues du secteur archivistique et, ou liées aux grands conflits du XX^{e} siècle, afin d'accroître le nombre d'entités pouvant être désambiguïsées à l'aide de référentiels externes. Enfin, les noms de personnes peuvent concerner des personnes peu connues, qui partagent par exemple la caractéristique d'avoir été interviewées par des chercheurs du CegeSoma pour décrire la vie quotidienne en temps de guerre : le Cegesoma est alors seul responsable de maintenir des autorités suffisamment complètes pour désambiguïser ces entités et éviter des problèmes de doublons à l'avenir.

4.2.4 Enrichissement

Une fois cette étape de réconciliation effectuée, il devient possible d'enrichir les autorités du CegeSoma grâce aux données contenues dans Wikidata [70].

69. Comme le relèvent Casalini *et al.*, « the amount of data needed for disambiguation varies with the *commonness* of the name » (Casalini *et al.*, 2018, p. 18).

70. Bien que cette thèse ne souhaite pas s'appesantir sur le sujet tant que la pérennité des notices d'autorité du Centre n'est pas encore assurée, ce procédé est également envisagé en sens inverse : lorsque le CegeSoma possède des informations qui ne sont pas encore présentes dans Wikidata (par exemple des pseudonymes, l'appartenance à un réseau de résistance, le fait que le CegeSoma possède des archives au sujet de cette personne, etc.), il pourrait les

En effet, lorsqu'un nom de personne a été associé à un identifiant Wikidata, il devient aisé de récupérer en masse d'autres informations stockées sur cette base de connaissance grâce à l'API de Wikidata [71]. Cette étape, qui peut être réalisée en quelques clics [72] à l'aide d'un service dédié du logiciel OpenRefine [73], permet de compléter des données parfois lacunaires, que le personnel n'a pas le temps d'étoffer lui-même. Ce type d'enrichissement est basé sur deux types de contenus [74] : premièrement, des identifiants externes tels qu'un identifiant SNAC ou un identifiant ISNI ; deuxièmement des déclarations associées à une personne, telle que sa date de naissance ou son lieu de décès, son genre ou encore un grade militaire. Cela pourrait par exemple être particulièrement utile dans le cadre des noms de personnes issus du thésaurus Pallas, étant donné que seules les années de naissance et de décès sont précisées, sans que ne figurent le jour et le mois en particulier.

Bien entendu une telle démarche ne peut être envisagée sans que ne se pose la question de la qualité des données, de leur fiabilité et de leur provenance : quid de la crédibilité du CegeSoma si des informations erronnées étaient diffusées sur une page estampillée CegeSoma ? Il s'agit de procéder à un arbitrage de type coûts-bénéfices afin de distinguer si les bénéfices qui peuvent en être tirés excèdent les risques que prendrait ainsi l'institution. Sachant que le personnel du CegeSoma connaît des situations de sous-effectif et que cette tendance n'a cessé de se renforcer au cours des dernières années, que nul n'a pour fonction principale de vérifier et de compléter les données d'autorité du Centre et que ces dernières s'avèrent particulièrement sommaires, il est apparu, en concertation avec la responsable du service de

enrichir de façon automatisée, sans devoir encoder chaque information manuellement. Par exemple, en juillet 2019, il n'existait pas encore de fiche Wikidata pour le résistant belge Albert Mélo, alors que le CegeSoma possède de nombreuses informations à son sujet (voir par exemple la page qui lui est dédiée sur Belgium WWII : https://www.belgiumwwii.be/destins-de-guerre/albert-melot.html).

71. https://www.wikidata.org/w/api.php.

72. En revanche, cela peut prendre plusieurs heures, en fonction du volume de données à rapatrier.

73. Étant donné l'existence de multiples tutoriels à ce sujet, nous ne détaillons pas ici l'intégralité du processus. Voir par exemple :

- ce tutoriel de Karen Li-Lun Hwang (https://medium.com/the-bytegeist-blog/enriching-reconciled-data-with-openrefine-89b885dcadbb)
- ce tutoriel de Mathieu Saby (https://msaby.gitlab.io/formation-openrefine-Lyon20191122/enrichir-ses-données-et-aligner-sur-des-sources-extérieures.html#réconcilier-des-données-avec-wikidata)
- ce tutoriel de la Map & Data Library (MDL) de l'université de Toronto (https://mdl.library.utoronto.ca/technology/tutorials/openrefine-augmenting-activity-2).

74. Nous pourrions également songer à un usage plus expérimental inspiré des travaux de Hwang, Karen Li-Lun (2017) : à savoir récupérer des *tags* de type indexation-matière, en partant des catégories Wikipédia associées à une personne, comme par exemple *Déporté résistant* (https://fr.wikipedia.org/wiki/Catégorie:Déporté_résistant).

numérisation, qu'un tel enrichissement automatisé pourrait être bénéfique à l'institution et à ses différents publics, qui profiteraient ainsi de davantage d'éléments contextuels.

En revanche, comme l'a souligné Garmendia (2019), il faut pouvoir *acknowledging in a transparent manner that data is imperfect and embracing uncertainty*. Cela se traduit ici par le fait de spécifier la source d'où provient l'information : Wikidata. Par chance, le logiciel Wikibase permet d'indiquer une référence au niveau de l'élément. Cela signifie qu'il est possible de prévoir par exemple d'indiquer comme source l'identifiant Wikidata correspondant, accompagnée de la « date de consultation », à savoir la date où l'enrichissement des données a été effectué. Si l'utilisateur veut connaître la provenance exacte de l'information ajoutée sur Wikidata, il pourra trouver l'information sur la fiche Wikidata correspondante, dans le cas où une référence a été ajoutée [75].

Si ce processus a été testé et est indéniablement prometteur dans le sens où il permettrait au CegeSoma d'enrichir massivement ses données tout en affichant clairement que certaines informations proviennent de sources externes qui ne relèvent pas de l'expertise du personnel, il faut toutefois noter que cela reste assez laborieux. En effet, cette démarche implique de nombreux aller-retours et adaptations entre le jeux de données préexistant, les propriétés et valeurs issues de Wikidata, et les propriétés et valeurs destinées à être ajoutées sur la Wikibase. Nous reviendrons en détail sur les questions de types de données prises en charge par Wikibase au cours de la section suivante, mais il faut savoir que cela a un impact sur ce type de processus d'enrichissement.

Imaginons par exemple extraire de Wikidata tous les lieux de naissance et de décès de plusieurs centaines de résistants belges, incluant Q3371461|Paul Henry de la Lindi [76]. Il s'avère que la fiche Wikidata de ce dernier contient des informations sur son lieu de naissance et de décès. Apprenant que son lieu de naissance est Q83407|Mons [77], nous avons le choix : nous pouvons soit récupérer le libellé dans la langue désirée, soit récupérer l'URI du concept (http://www.wikidata.org/entity/Q83407). Cependant, étant donné que, comme explicité en début de section, nous avons opté pour un référentiel maintenu localement pour les noms de lieux situés en Belgique, il faudra commencer par réconcilier ce libellé ou cet URI à l'entité corres-

75. Un parti pris plus strict pourrait être de ne réutiliser des informations issues de Wikidata seulement lorsqu'une référence externe a été associée (sur Wikidata) à la déclaration concernée, mais l'implémentation serait dès lors plus laborieuse et le volume de données ajoutées moindre.

76. https://www.wikidata.org/wiki/Q3371461.

77. https://www.wikidata.org/wiki/Q83407.

pondante du référentiel[78], ce qui alourdit considérablement la procédure. En outre, il faut garder à l'esprit que ces données ne seront pas synchronisées et ne bénéficieront pas des éventuelles corrections qui pourraient être effectuées à la source, directement sur Wikidata.

Une alternative, qui sera présentée en détail au cours du chapitre 5, consiste à laisser l'information là où elle est et ainsi éviter de la dupliquer[79], en allant seulement la récupérer au moment voulu, à la demande. Dans notre cas précis, il s'agit de faire appel aux possibilités offertes par les requêtes SPARQL fédérées, permettant d'interroger simultanément plusieurs bases de connaissance – à savoir : notre propre instance Wikibase et Wikidata. Ainsi, au lieu de mélanger sur notre Wikibase des informations issues du CegeSoma et des informations issues de sources externes – avec les risques de confusion et d'information obsolète que cela suppose –, il s'agit d'accoler aux résultats d'une requête des informations provenant d'autres sources.

4.3 Modélisation

Cette section détaille la façon dont les données relatives aux personnes ont été modélisées. Ces modèles, co-créés en collaboration avec des membres du personnel scientifique du CegeSoma[80], ambitionnent de couvrir l'intégralité des cas de figure rencontrés dans le cadre des données du CegeSoma. À l'instar de l'ontologie Records in Contexts (International Council on Archives Expert Group on Archival Description, 2019a), le cadre de description proposé ici a pour ambition d'être flexible : il doit permettre d'accueillir des données d'autorité archivistiques dont le niveau de description varie énormément. Cela signifie que l'ensemble des propriétés n'a pas vocation à être utilisé de façon systématique, mais qu'à chaque besoin doit correspondre une méthode standardisée de description de l'information. En outre, il faut noter que ce cadre de description représente un canevas de départ, suceptible

78. Il s'agira donc, *in fine*, d'effectuer un mapping entre l'élément Wikidata et l'élément correspondant de notre Wikibase.

79. Ce qui rejoint les recommandations formulées par le gestionnaire des bases de données des Archives de l'État, mais également par l'industrie : « The client should need to input only the private and confidential knowledge or any knowledge that the system does not yet know. Isolation, federation, and online updates of the base and domain layers are some of the major issues that surface because of this requirement » Noy *et al.* (2019).

80. Comme le relèvent Lovins et Hillmann, « technologists are not always accustomed to dealing with maintainers, but they need to work with those who fully understand the data, and not insist that maintainers understand the full technology stack. » (Lovins et Hillmann, 2017).

d'être adapté et complété à l'avenir, en fonction des réalités et des besoins rencontrés [81].

Concrètement, outre le libellé, la description et les éventuels alias, déjà présentés au cours du chapitre 2 et constituant le tronc commun des éléments Wikibase [82], le modèle de données Wikibase requiert d'établir des propriétés auxquelles sera associé un type de valeur, comme l'illustrera en détail la première sous-section dédiée aux données d'identification. Ces propriétés pouvant être créées librement donnent la possibilité aux institutions de dépasser les limites propres aux logiciels traditionnels, sans devoir toutefois passer par d'importants coûts de développement pour développer leur propre outil. Elles peuvent ainsi s'émanciper de logiciels prescrivant une certaine vision du monde à l'aide de modèles préétablis et pouvant être générateurs de frustrations (van Hooland et Verborgh, 2014). Cependant, il faut garder à l'esprit que les modèles sur mesure nécessitent d'énièmes alignements et sont exigeants à maintenir (Smith-Yoshimura, 2018a; Lovins et Hillmann, 2017). Pour cette raison et conformément aux recommandations du W3C (W3C, 2014), le modèle de données de cette Wikibase est donc destiné à être conçu, dans la mesure du possible, à partir de propriétés préexistantes – dans l'idéal des propriétés Wikidata [83] – afin de favoriser l'interopérabilité des données.

Comme nous le verrons en fin de section, la réutilisation de propriétés préexistantes [84] n'a pas été systématiquement possible. Or, l'élaboration d'un modèle de données est toujours le résultat d'un ensemble de choix qui peuvent se révéler d'une grande complexité. En effet, la *mise en propriétés* d'informations parfois pointues s'avère un exercice délicat, par exemple lorsqu'il s'agit de personnes condamnées pour faits de collaborations : l'étape de modélisation des peines et procès requiert d'anticiper en envisageant les différents cas de figure susceptibles d'être ajoutés ultérieurement, tout en réfléchissant à la façon dont de telles informations pourront ensuite être interrogées. Cette section vise à présenter à l'aide d'exemples ces propriétés et les réflexions sous-jacentes à leur création : premièrement les propriétés plus « génériques » relatives à des données d'identification, deuxièmement les propriétés relatives à la Seconde Guerre mondiale, troisièmement les propriétés concernant des relations à d'autres personnes ou ressources. En-

81. Comme le relève Boydens, « dans le cadre d'un système d'information empirique, un objet n'est jamais nécessairement identique à lui-même. Nous entendons par là, d'une part, qu'à un même concept empirique peut correspondre à un instant donné une pluralité de significations interagissantes et, d'autre part, qu'interagissant, ces signfications sont évolutives » (Boydens, 2001).

82. Voir : https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

83. Comme l'ont souligné Zhou *et al.*, Wikidata est en effet considérée comme une base de connaissance utile dans le cadre de tâches d'alignement (Zhou *et al.*, 2020, p. 3).

84. Le chapitre qui suit détaille la manière dont ces propriétés sont concrètement créées.

fin, une dernière sous-section s'intéresse à leur alignement avec Wikidata et l'ontologie Records in Contexts.

4.3.1 Données d'identification

Cette sous-section vise à présenter les propriétés et éléments requis pour intégrer les informations permettant d'identifier une entité. La figure 4.4 montre un schéma du modèle utilisé, avec ici un exemple décrivant la résistante belge Andrée de Jongh. L'entité Wikibase Q10|Andrée de Jongh [85] est caractérisée à l'aide de différentes propriétés (en rose), dont chacune est identifiée à l'aide d'un libellé, par le biais de la propriété *rdfs :label* ; par exemple, la propriété P31 [86] correspond en français à la date de naissance, dans le contexte de notre Wikibase. À chaque propriété est ensuite associée une ou plusieurs valeurs [87] (en blanc, dans les rectangles bleus), mais attention, cette valeur doit correspondre à un certain type de données [88]. Ainsi, la propriété P31|date de naissance [89] n'accepte par exemple que des valeurs de *type : time*, à savoir des dates issues du calendrier géorgien ou julien, stockées sous forme d'horodatage inspiré de la norme ISO 8601, comprenant entre 4 et 16 chiffres (selon le degré de précision). Cela implique qu'une date approximative exprimée sous forme de texte dans les données d'origine, comme par exemple *entre 1940 et 1942*, devra d'abord faire l'objet d'une transformation.

85. https://adochs.arch.be/wiki/Item:Q10

86. Nous attirons l'attention sur le fait que les éléments et propriétés Wikibase sont désignés par une graphie similaire à celle des éléments et propriétés Wikidata (Qxy et Pxz).

87. Comme nous l'avons vu au cours du chapitre 2, cette particularité du modèle Wikibase permet notamment de faire cohabiter des faits contradictoires.

88. Pour des détails sur le modèle de données Wikibase, déjà évoqué au cours du chapitre 2, voir : https://www.wikidata.org/wiki/Wikidata:Glossary/fr ; https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer ; https://www.wikidata.org/wiki/Help:Data_type/fr.

89. https://adochs.arch.be/wiki/Property:P31.

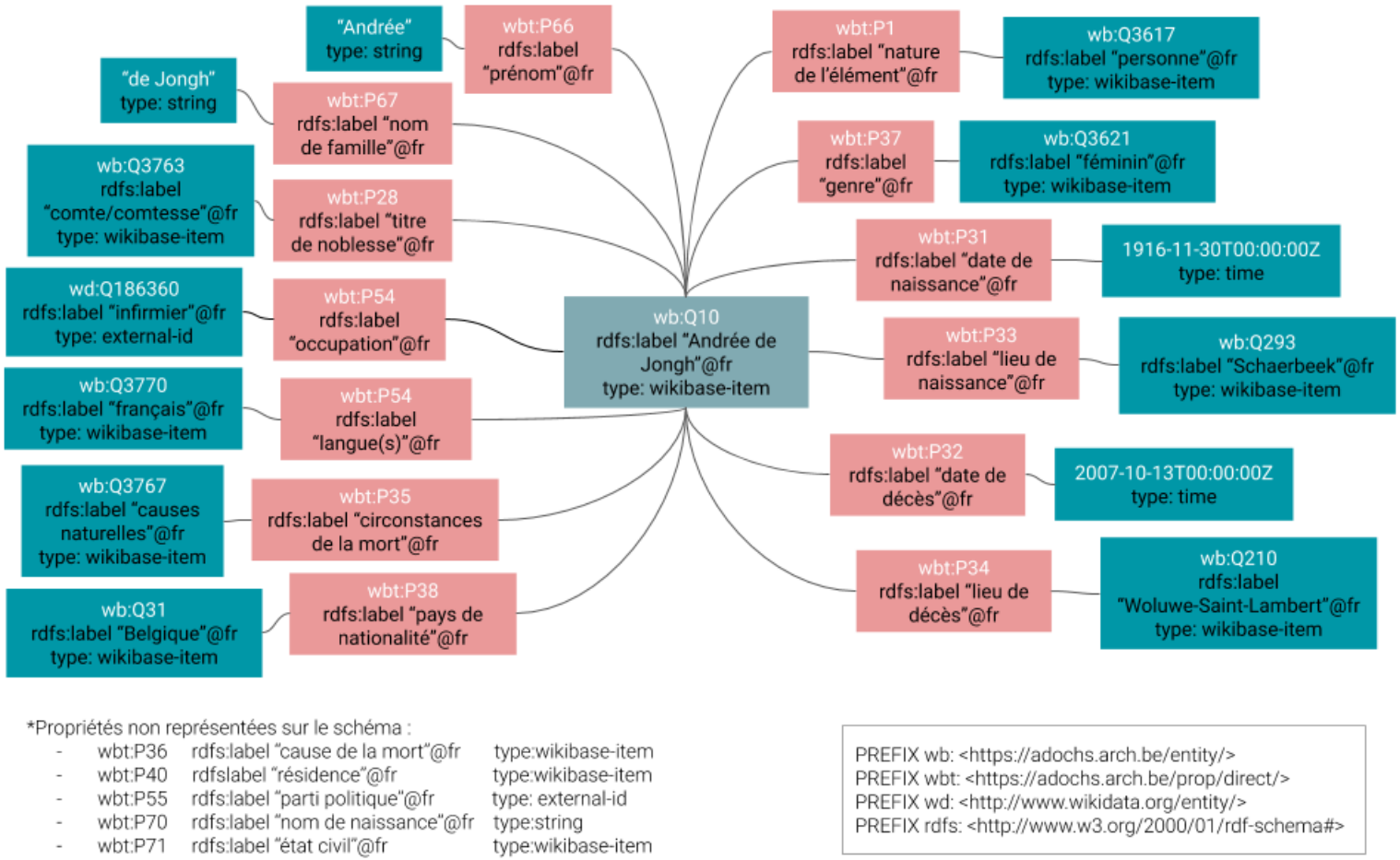


FIGURE 4.4 – Schéma du modèle de données utilisé pour encoder des informations d'identification d'une personne au sein de la Wikibase ; exemple portant sur la résistante belge Andrée de Jongh|Q10.

Il en va de même pour un nom de lieu : dans le cas des données du Cege-Soma, les noms de lieux sont systématiquement exprimés en langage naturel dans le fichier d'origine, sous la forme d'une chaîne de caractères. Or, l'un des principaux intérêts d'une instance Wikibase réside précisément dans le fait de pouvoir tirer parti de la puissance de données structurées lisibles par une machine : il paraît donc plus stratégique que le type de valeur attendu corresponde à un élément de la Wikibase [90], permettant d'identifier un lieu de manière non ambiguë à l'aide d'un URI. Cela signifie qu'un effort supplémentaire sera requis au moment de la création des données : d'une part, il faudra commencer par créer les éléments Wikibase correspondant – à savoir des noms de lieux –, d'autre part, si l'encodage se fait initialement dans un fichier local, il faudra réconcilier les noms de lieux à l'URI correspondant, ce qui relève du traitement de la *dette sémantique*. La création des propriétés et le choix du type de données associé à chaque propriété constituent donc une étape cruciale, qui a une incidence tant sur le processus de création et d'importation des données que sur la manière dont ces dernières pourront ensuite être interrogées.

Dans le cadre des données d'identification, la majorité des propriétés [91] acceptent comme valeur un élément Wikibase ; deux propriétés [92] acceptent comme valeur des dates ; trois autres [93] acceptent comme valeur des chaînes de caractères et enfin, deux propriétés [94] acceptent comme valeur des identifiants externes. Comme souligné ci-dessus, ces choix de modélisation sont stratégiques et méritent un arbitrage de type coûts-bénéfices. Nous détaillons ici ces raisonnements pour trois cas de figure : les propriétés liées aux noms ; les propriétés liées aux lieux ; les propriétés liées à des occupations et à des partis politiques.

Premièrement, en ce qui concerne les noms de personne : au départ, il nous semblait préférable d'utiliser des éléments plutôt que des chaînes de caractères, à l'instar de ce que le créateur de l'instance FactGrid a choisi de faire a posteriori [95], de manière à pouvoir mieux tirer partie des requêtes SPARQL [96]. Ce choix semble en effet le plus prometteur au niveau de l'ex-

90. *Type : wikibase-item*.

91. À savoir : P1|nature de l'élément ; P28|titre de noblesse ; P33|lieu de naissance ; P34|lieu de décès ; P35|circonstances de la mort ; P36|cause de la mort ; P37|genre ; P38|pays de nationalité ; P40|résidence ; P54|langue ; P71|état civil.

92. P31|date de naissance ; P32|date de décès.

93. P66|prénom ; P67|nom de famille ; P70|nom de naissance.

94. P54|occupation ; P55|parti politique.

95. Passant ainsi d'une seule chaîne de caractères incluant de façon indistincte trois prénoms *(Johann Joachim Christoph)* à l'utilisation de prénoms utilisés sous la forme d'éléments Wikibase auxquels peuvent être adjoints un qualificatif précisant leur ordre, voir par exemple Q133|Johann Joachim Christoph Bode (https://database.factgrid.de/wiki/Item:Q133).

96. Échange privé sur Twitter avec Olaf Simons, 22 mai 2019.

ploitation des données [97] et de l'encodage de nouvelles informations au sein de l'interface graphique – facilitée par la fonctionnalité d'autocomplétion lorsque les noms sont sélectionnés parmi les noms contenus dans la Wikibase.

Cependant, dans le cadre du CegeSoma, les contacts avec le terrain nous ont permis de poser divers constats : en premier lieu, la présentation de l'instance et de son fonctionnement auprès du personnel a déjà suscité un certain nombre de questions et de craintes quant à sa prise en main par des bénévoles parfois âgés, avant même que ce genre de scénario n'ait été évoqué, laissant penser que cette couche de complexité supplémentaire [98] pourrait compromettre son adoption ; deuxièmement, il semblerait d'après nos échanges avec le personnel que la majorité des données nominatives soient appelées à l'avenir à continuer à être encodées dans des fichiers Excel avant d'être importées dans la Wikibase, rendant secondaire l'argument de l'autocomplétion et signifiant soit un véritable ralentissement lors de l'encodage – si chaque prénom, nom de famille ou nom de naissance doit d'abord être sélectionné dans une liste potentiellement incomplète –, soit un important travail de réconciliation *a posteriori* – si les noms sont encodés tels que rencontrés sur le document dans un premier temps –, avec toutes les difficultés et les questionnements que peut générer la présence de graphies légèrement différentes ; troisièmement, cela vient alourdir la prise en charge des jeux de données préexistants en accroissant significativement les étapes de pré-traitement et de mise en correspondance.

Sachant cela, la plus-value représentée par l'utilisation d'éléments issues de la Wikibase plutôt que des chaînes de caractères ne semble pas, dans le cas présent, justifier les efforts supplémentaires que cela requiérerait – dans un premier temps du moins [99]. Par ailleurs, la possibilité existe de pouvoir

97. FactGrid donne par exemple l'opportunité à ses utilisateurs de retrouver très facilement toutes les personnes portant un certain nom de famille, exemple pour *Möller* : https://database.factgrid.de/wiki/Special:WhatLinksHere/Item:Q25777.

98. Dans le cas où les données sont d'abord encodées dans un fichier intermédiaire avant d'être importées dans la Wikibase, utiliser un URI plutôt qu'une chaîne de caractère ajoute de la complexité.

99. Notons qu'il pourrait également être envisagé de faire temporairement cohabiter les deux. Wikidata a par exemple opté pour cette solution dans le cadre des noms d'auteurs de plusieurs centaines de milliers de publications scientifiques – destinées à pouvoir être citées dans des articles Wikipedia – : P2093|auteur (chaîne) vient suppléer P50|auteur lorsque seul le nom de l'auteur est connu, sans autre information permettant de le distinguer d'homonymes. Dans ce cas, seule une chaîne de caractères est dès lors utilisée plutôt qu'un élément Wikidata à part entière. Il est toutefois précisé qu'il s'agit avant tout d'une solution de contournement et d'un compromis pragmatique (Wikidata, 2020a).

L'ontologie Records in Contexts prévoit également ce genre de cas de figure : elle combine des *datatype properties* – telles que rico :name – acceptant comme valeur des chaînes de caractères, à des classes permettant de traiter les noms comme des entités à part entière – comme par exemple rico :AgentName. Cette coexistence est justifiée par la volonté d'offrir

formuler des requêtes à l'aide de la commande *FILTER*, afin de rechercher une variable composée d'un littéral [100], bien que cette solution ne soit pas optimale car plus gourmande en ressources [101].

Deuxièmement, les entités géographiques constituent un élément crucial : permettant de documenter lieux de naissance, de décès, de résidence ou encore le pays de nationalité, elles font partie des informations déterminantes lorsqu'il s'agit de distinguer des personnes portant le même anthroponyme. L'utilisation de valeurs stockées sous la forme de chaînes de caractères nous semblait ici largement contre-productive et n'a pas été envisagée plus avant. En ce qui concerne les localités, comme par exemple les communes belges, nous avons déjà évoqué auparavant l'intérêt de travailler avec le référentiel utilisé par les Archives de l'État pour tout ce qui concerne le territoire belge. Pour l'instant, ces entités ont toutes été importées dans la Wikibase après avoir été croisées avec des données issues de l'Institut national de Statistique de Belgique et des données tirées de Wikidata. Il sera toutefois nécessaire à l'avenir d'implémenter un système de synchronisation entre la base de données des Archives de l'État et la Wikibase, afin que les données puissent rester complètes et à jour [102].

La question est cependant plus large que cela et doit également être posée pour des lieux situés à l'étranger. C'est une question délicate, qui ne peut être ignorée, mais qui n'a en revanche pas encore été tranchée. Cela s'explique par le fait que la majorité des lieux présents dans notre échantillon sont situés en Belgique et qu'une conciliation est en cours avec les Archives

une ontologie immédiatement utilisable par tous, notamment par ceux qui voudraient créer des jeux de données de données RDF à partir de leurs métadonnées archivistiques sans être en mesure – dans un premier temps du moins – d'utiliser systématiquement des URIs (International Council on Archives Expert Group on Archival Description, 2019a). Cette coexistence ne paraît cependant pas utile pour l'instant dans le cadre du CegeSoma, étant donné que l'intégralité des prénoms et noms rencontrés sont constitués de chaînes de caractères.

100. Par exemple en le combinant à des fontions comme *STRSTARTS*, *STRENDS*, *CONTAINS* ou *REGEX*, voir : https://www.w3.org/TR/sparql11-query/#func-strings.

101. Comme l'explique l'utilisateur Wikidata ArthurPSmith – qui développe et maintient un certain nombre d'outils de cette base de connaissance – : « Traditionally graph databases index URI's but not string (or partial-string) values, so item-based matches can be very efficient, while string matches require fetching a lot of data and doing the string comparison on that, [and are therefore] much more resource intensive », [Message Telegram] Arthur Smith, 16.08.20.

Ainsi, une requête rudimentaire lancée sur le point d'accès SPARQL de notre instance Wikibase révèle une légère différence de temps de calcul : une recherche de personne basée sur un nom de famille composé d'un élément Wikibase affiche un résultat après 179 millisecondes, tandis qu'une requête de personne basée sur un nom de famille contenant une certaine chaîne de caractère affiche le même résultat après 204 millisecondes : une différence susceptible d'être plus significative et contraignante si l'instance, contenant aujourd'hui moins de 10 000 éléments, continue de croître.

102. Ce questionnement est en suspend étant donné que les Archives de l'État sont actuellement en réflexion quant à la manière de publier et pérenniser l'accès à ces entités géographiques, qui ne sont pour l'instant uniquement disponibles en interne.

de l'État afin de favoriser des pratiques homogènes – l'ensemble des dépôts des Archives de l'État étant confronté à ce même type de questionnement. Une analyse approfondie est par ailleurs requise afin de pouvoir déterminer quels référentiels seraient les plus à même de répondre aux besoins de l'institution au niveau de l'évolution des noms de lieux dans le temps et l'espace, ce qui va au-delà de notre objet d'étude [103].

Une fois ce choix posé, il sera encore nécessaire de déterminer sous quelle forme ce référentiel sera utilisé : soit il sera directement importé dans l'instance Wikibase – sous la forme de nouveaux éléments décrits à l'aide de diverses déclarations permettant d'intégrer des données comme des coordonnées géographiques –, soit les références aux lieux pourront se faire à l'aide d'identifiants externes. Là où la première option implique de maintenir la liste de données localement, avec tous les challenges de multilinguisme, de synchronisation et de mise à jour de l'information que cela implique, la seconde option permet d'éviter ces efforts de maintenance, mais se traduit par une moins grande autonomie quant à la qualité des données et un format beaucoup plus limité dans le cadre de la Wikibase [104].

Ce choix se télescope par ailleurs avec les données préexistantes : un lieu de naissance situé en Belgique est actuellement indiqué à l'aide de la propriété P33|lieu de naissance, qui accepte comme valeurs des éléments de la Wikibase. Utiliser des noms de lieux issus de référentiels externes tels que GeoNames sous forme d'identifiants externes reviendrait donc à devoir créer deux propriétés distinctes : la première pour un lieu de naissance situé en Belgique, dont le type de données serait un élément Wikibase ; la seconde pour un lieu de naissance situé à l'étranger, dont le type de données serait un identifiant externe. Une telle configuration ne semble pas heureuse sachant que ce processus devra être répété pour tous les types de lieux (lieux de décès, de domicile, d'arrestation, etc.). Une variante pourrait consister à conserver la même propriété et à utiliser comme valeur un élément comme « lieu à l'étranger », dont le lien vers l'autorité issu d'un référentiel externe serait précisé en tant que qualificateur. Enfin, une solution intermédiaire, qui serait à même de dépasser ces difficultés, réside dans le projet de fédération d'instances Wikibase sur lequel travaille actuellement l'équipe de

103. À titre indicatif, nous présentons toutefois quelques premières investigations relatives aux noms de pays, dans l'Annexe 5, page 345.

104. Par exemple, un nouveau nom de lieu ne pourra pas être encodé directement au sein de l'interface avec les fonctionnalités d'autocomplétion permises par Wikibase, et des informations complémentaires telles que des formes alternatives du nom, des dates d'existence ou des coordonnées géographiques ne seront pas stockées localement, elles seront seulement (potentiellement) présentes au sein des données d'origine, rendant par exemple beaucoup plus laborieuse la visualisation de données sur une carte, car nécessitant l'utilisation de requêtes fédérées, gourmandes en ressource et augmentant considérablement la verbosité des requêtes SPARQL.

développement de Wikidata et de Wikibase (Wikidata, 2020b). Avec un tel système, les données issues de Wikidata pourraient alors être réutilisées, tout en continuant à être actualisées et potentiellement enrichies localement à l'aide d'identifiants internes ou d'autres jeux de données.

Troisièmement, il existe des propriétés qui ne représentent pas le cœur de métier du CegeSoma et de ses données. C'est le cas par exemple de l'appartenance à un parti politique : il est possible que cette information soit connue pour une minorité des personnes liées aux collections du CegeSoma, mais ce n'est pas du tout systématique. De plus, quand bien même l'information est connue, elle est exprimée en langage naturel, l'institution – dont les ressources sont limitées, comme nous l'avons vu – n'ayant pas comme priorité de maintenir un vocabulaire contrôlé reprenant les différents partis politiques ayant existé au XX^{e} siècle en Belgique et dans le reste de l'Europe. Il en va de même pour les professions : s'il arrive que cette information soit connue, elle s'avère être encodée sous forme de texte libre, tantôt en français, tantôt en néerlandais, et non pas à l'aide d'un vocabulaire contrôlé.

Dans ce genre de cas, il semble utile de tirer parti des liens de proximité avec la base de connaissance Wikidata pour réutiliser les données qu'elle contient. En effet, une base généraliste contenant des informations sur plus de six millions de personnes (Wikidata, 2020b) est appelée à être fréquemment confrontée à ce type d'informations – professions ; appartenance à un parti politique. La stratégie ici vise donc à utiliser pour chaque profession ou parti politique une valeur composée de l'identifiant numérique Wikidata correspondant : par exemple pour une personne dont la profession était menuisier, c'est l'élément Wikidata Q326358 [105] qui sera stocké dans la Wikibase (sous forme d'identifiant externe) [106]. De même, pour une personne qui a été membre du parti d'extrême-droite Rex : c'est l'élément Wikidata Q314493|Rex [107] qui sera encodé comme valeur. Un tel mécanisme ne va cependant pas sans inconvénient : il crée une rupture dans le processus d'encodage des données en condamnant l'utilisateur à devoir consulter Wikidata afin de trouver l'entité correspondante [108] ; de plus il engendre un risque de confusion entre Wikidata et Wikibase pour les utilisateurs qui ne seraient

105. https://www.wikidata.org/wiki/Q326358.

106. Nous pourrions également envisager d'importer toutes ces données directement dans notre Wikibase, cependant, cela nous priverait de la mise à jour des données ou nous contraindrait à d'importantes contraintes de synchronisation (et multiplierait inutilement les étapes de mise en correspondance, en cas de désir de réinjecter certaines informations dans Wikidata). De plus, cela générerait très probablement du bruit dans les recherches, sachant que plus de 300 000 éléments Wikidata appartiennent à la classe https://w.wiki/ZdZ.

107. https://www.wikidata.org/wiki/Q314493.

108. Cela pourrait toutefois être évité en passant par exemple par le logiciel OpenRefine, qui permettrait de réconcilier les données à la fois avec Wikidata et la Wikibase concernée, sans devoir changer de plateforme.

pas familiers avec l'infrastructure et le modèle de données utilisé ; en outre, il s'accompagne d'une certaine opacité des données : indiquer comme valeur un identifiant externe tel que *Q314493* ne peut rivaliser avec un élément doté d'un libellé explicite [109] ; par ailleurs, subsiste la question de la qualité des données : se reposer sur Wikidata permet certes de ne pas devoir prendre en charge la maintenance des données, mais cela signifie également ne pas avoir un contrôle complet sur la qualité et être potentiellement confronté à des difficultés comme des problèmes de bruit [110], d'incomplétude [111], d'incohérence [112], ou encore de manque de pertinence [113] ou de manque de représentativité [114].

Finalement, bien que Wikidata soit librement éditable et qu'il soit dès lors possible d'imaginer un workflow incluant tant l'amélioration d'entités Wikidata existantes que l'éventuelle création de nouvelles entités, il faut noter que cette question de la qualité renvoie à celle, plus large des *sources authentiques* à privilégier : Wikidata se prête-t-elle à ce rôle de référentiel,

109. L'instance Lingua Libre l'a bien compris et a développé à cette fin un script permettant de remplacer les identifiants externes Wikidata par le libellé et la description de l'élément correspondant, dans la langue de l'interface, voir : https://lingualibre.org/wiki/MediaWiki:Common.js.

110. Par exemple, lorsque nous avons effectué une réconciliation semi-automatisée des professions listées dans un fichier du CegeSoma, l'entrée *étudiant* a été erronément réconciliée avec l'entité Wikidata Q2248623|étudiant (https://www.wikidata.org/wiki/Q2248623) à la place de l'entité plus générique Q48282|étudiant (https://www.wikidata.org/wiki/Q48282), comme permet de le comprendre leur description respective : « au Moyen Âge, clerc ou élève d'une université » pour la première, « personne intégrée dans un cursus scolaire pour la seconde ». Ainsi, l'importante quantité de professions documentées par Wikidata se traduit par un risque accru d'imprécision, mais aussi potentiellement par un certain manque de cohérence, en fonction de l'interprétation des données par la personne en charge de l'encodage ou de la réconciliation.

111. Par exemple, la description de l'entité Q314493|Rex (https://www.wikidata.org/wiki/Q314493) n'est pour l'instant pas disponible en allemand, tandis que Q3276580|magasinier (https://www.wikidata.org/wiki/Q3276580) est libellé et décrit uniquement en français !

112. Par exemple, le niveau de granularité des descriptions varie énormément en fonction des éléments, mais également entre les traductions d'un même élément (ainsi, Q627325|graphiste (https://www.wikidata.org/wiki/Q627325) est décrit sommairement en allemand – *Beruf* – tandis qu'il est défini plus minutieusement en français – professionnel de la communication qui conçoit des solutions de communication visuelle). On constate également des acceptions plus ou moins larges d'une même profession selon la langue de description (par exemple, Q185351|juriste (https://www.wikidata.org/wiki/Q185351), décrit en anglais commme *legal scholar or academic, a professional who studies, teaches, and develops law*, là où la description en néerlandais indique plus sobrement : « iemand die het recht bestudeert »). Par ailleurs, il arrive qu'un élément recouvre des réalités différentes selon la langue consultée (ainsi l'élément Q185196|sage-femme (https://www.wikidata.org/wiki/) et l'élément Q13638192|obstétricien (https://www.wikidata.org/wiki/Q13638192) sont tous les deux désignés indistinctivement par le terme *ostetrica* en italien, alors qu'ils correspondent à des réalités différentes, en français du moins).

113. Ainsi, l'élément Wikidata Q484188|tueur en série (https://www.wikidata.org/wiki/Q484188) s'avère être considéré comme un métier (P31|nature d'élément) !

114. Nous avons par exemple rencontré des difficultés pour retrouver sur Wikidata certains métiers tombés en désuétude comme receveur de tram.

ou serait-il préférable de cibler un référentiel spécialisé pour chaque cas de figure rencontré ? Par exemple, dans le contexte des professions, une alternative serait d'utiliser la classification ISCO-08 (International Standard Classification of Occupations), émise en 2008 par l'Organisation Internationale du Travail (OIT), reprenant plus de 7 000 professions et utilisée depuis 2011 par l'institut national de statistiques de Belgique [115] ou, mieux encore, ESCO (European Skills, Competences, Qualifications and Occupations), le vocabulaire contrôlé de la Commission européenne pour les professions, lancé en 2017, incluant près de 3 000 professions traduites dans 27 langues [116] et qui est par ailleurs adapté au Web sémantique, étant donné que chaque profession est identifiée de façon pérenne à l'aide d'un URI [117]. La question se pose donc de savoir s'il serait préférable de favoriser ce type de référentiel spécialisé, quitte à multiplier les sources et ajouter de la complexité dans les instructions à donner aux personnes chargées de l'encodage des données ? Une solution plus pragmatique serait sans doute que Wikidata systématise son utilisation de liens d'équivalence vers la classification ISCO-08 [118] ou vers la classification ESCO [119]. Cela permettrait de capitaliser les liens entre Wikidata et Wikibase – avec toutes les possibilités d'enrichissement mutuel que cela permet –, tout en limitant les problèmes de qualité des données en passant par exemple par un sous-ensemble de données correspondant aux données ESCO.

Pour l'heure, la solution la plus adéquate semble donc de passer par un identifiant Wikidata stocké sous forme d'identifiant externe, en dépit des difficultés soulignées, tout en tentant de régler le problème d'opacité posé par ce type de données en utilisant un script permettant de remplacer l'identifiant unique par le libellé et la description de l'élément dans la langue de l'interface. La figure 4.5 montre un exemple de réutilisation de ce script : la profession de Andrée de Jongh est stockée sous forme d'identifiant Wikidata, mais affichée ici dans un format lisible par des humains.

115. https://statbel.fgov.be/fr/propos-de-statbel/methodologie/classifications/classification-internationale-type-des-professions.
116. https://ec.europa.eu/esco/portal/howtouse/21da6a9a-02d1-4533-8057-dea0a824a17a.
117. Exemple : ouvrier boucher http://data.europa.eu/esco/occupation/84dfc210-db07-4a83-87fe-d6131907ea83.
118. Ce lien d'équivalence existe déjà et peut exprimer via la propriété Wikidata P8283|code d'occupation ISCO-08 (https://www.wikidata.org/wiki/Property:P8283), cette dernière n'a cependant été utilisée qu'à quatre reprises (https://w.wiki/Zdu) jusqu'à présent !
119. La propriété P4652|identifiant ESCO d'une profession (https://www.wikidata.org/wiki/Property:P4652) existe, mais elle n'a été utilisée qu'à 249 reprises jusqu'à maintenant.

FIGURE 4.5 – Exemple d'utilisation de la propriété P54|occupation utilisant comme valeur un identifiant externe Wikidata ; exemple portant sur la résistante belge Andrée de Jongh|Q10.

4.3.2 Seconde Guerre mondiale

Avant de passer à cette vue exhaustive des données relatives à la Seconde Guerre mondiale, il faut commencer par souligner qu'une importante question préliminaire fut celle de la granularité : jusqu'à quel niveau de détail est-il pertinent d'aller ? Si la majorité des données préexistantes est relativement pauvre, faut-il se cantonner à cela, ou est-ce plus stratégique de pouvoir couvrir le plus de cas de figures susceptibles de se présenter, dans l'optique où cette Wikibase pourrait être utilisée non seulement au quotidien par le personnel occupé à décrire les collections, mais également par des chercheurs intéressés à publier les données issues de leurs travaux de recherches ?

Sachant qu'il n'existe pas de franche ligne de démarcation entre les données de la recherche et les données d'autorité utiles pour identifier un individu sans ambiguïté [120], la stratégie adoptée, en accord avec des experts du domaine travaillant au CegeSoma [121], fut de créer un ensemble de propriétés permettant de couvrir de façon exhaustive les différents champs des jeux de données utilisés, tout en prenant en compte quelques données en puissance, telles que les informations relatives aux personnes exécutées pour faits de collaboration, qui existent « virtuellement » mais n'ont pas encore été encodées de façon systématique dans un fichier.

En revanche, le nombre de propriétés a été limité au maximum afin d'éviter d'avoir des propriétés trop spécifiques susceptibles de n'être que peu utilisées, de créer de la confusion dans le chef de la personne responsable de l'encodage de nouvelles données et de rendre plus laborieuse la formu-

120. Ainsi, le fait de savoir qu'une personne a été impliquée dans un certain mouvement de résistance au cours de la Seconde Guerre mondiale peut présenter un fort potentiel discriminant dans le cadre du traitement d'entités nommées qui posséderaient des mêmes prénoms, noms et années de naissance et de décès.

121. Comme le relèvent Koho *et al.*, « the Linked Data approach requires tighter cooperation with the domain experts and data publishers, especially in the creation phase of historical information (Boonstra *et al.*, 2004), than more traditional data publishing ways » (Koho *et al.*, 2019a).

lation de requêtes. Ainsi, pour un certain nombre d'informations (comme des éléments de contexte tels que des dates ou des lieux), le recours à des qualificatifs permettant de préciser une propriété a été privilégié. En effet, conformément à cette recommandation adressée dans le cadre de la modélisation de données relatives à des procès de sorcellerie s'étant tenus en Écosse au début de l'époque moderne (Wikidata, 2020), plutôt que de créer des propriétés très spécifiques telles que *date d'arrestation*, il semble préférable d'utiliser au maximum une propriété générique comme *événement-clé*, à laquelle sera combiné un élément caractérisant l'événement (par exemple : *arrestation*), ainsi que des qualificatifs permettant de donner des informations sur le contexte (par exemple : *date et lieu d'arrestation*).

Par ailleurs, il faut relever ici qu'il est parfois nécessaire de faire preuve de créativité au cours du processus de modélisation des données, afin d'éviter de créer des propriétés stériles stockant uniquement des valeurs binaires de type *oui / non*. En effet, au cours de notre exploration préparatoire des données et de nos entretiens avec certains membres du personnel, nous avons rencontré à plusieurs reprises ce genre de données à usage limité ; par exemple : une demande de statut de reconnaissance a-t-elle été adressée pour cette personne auprès de la Sûreté de l'État à l'issue de la Seconde guerre mondiale, *oui* ou *non* ? Vu la structure en triplets (sujet - prédicat - objet) caractérisant le RDF, il semble plus puissant et utile de favoriser des objets contenant une information à haute valeur ajoutée. Ainsi, dans le cadre de l'introduction d'une demande de statut de reconnaissance, il semble beaucoup plus porteur d'opter pour un objet recelant une information supplémentaire, à savoir le type de statut demandé [122].

122. C'est-à-dire : résistant armé, agent de renseignements et d'actions, résistant par la presse clandestine, résistant civil ou encore prisonnier politique.

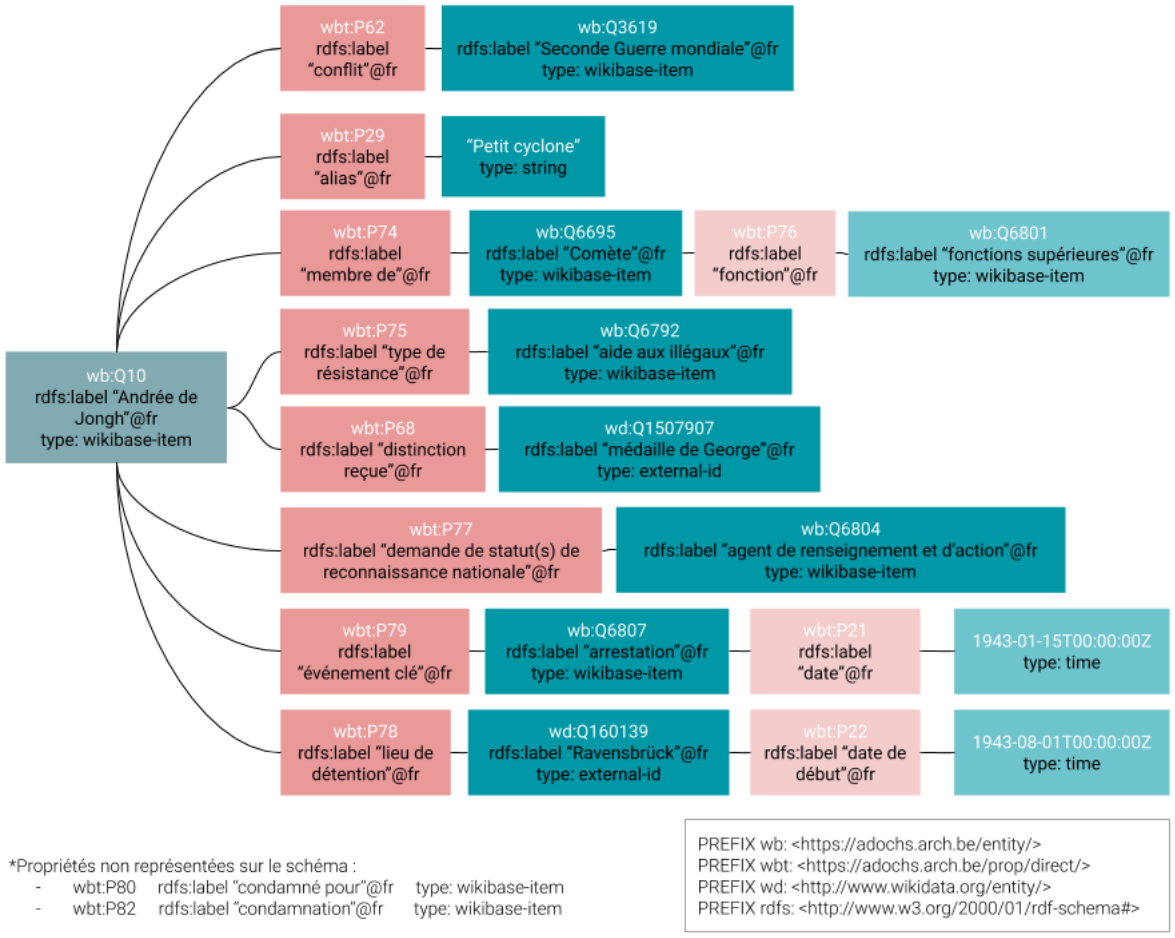


FIGURE 4.6 – Schéma du modèle de données utilisé pour saisir des informations relatives à la Seconde Guerre mondiale ; exemple portant sur la résistante belge Andrée de Jongh|Q10.

La figure 4.6 montre un schéma du modèle de données utilisé pour saisir des informations relatives à la Seconde Guerre mondiale, toujours avec l'exemple portant sur la résistante belge Andrée de Jongh. Il est construit sur le même principe que le schéma relatif aux données d'identification, à la différence qu'il prévoit également des qualificatifs accompagnés de leurs valeur respective, représentés dans une teinte légèrement plus claire. À nouveau, la plus grande partie des propriétés acceptent comme valeur un élément Wikibase [123] ; deux propriétés utilisées comme qualificatifs [124] attendent comme valeur des dates ; une propriété [125] accepte comme valeur des chaînes de caractères, et enfin, deux autres propriétés [126] sont basées sur des identifiants externes.

Il faut noter que ce travail de modélisation implique un important travail de standardisation de l'information, lorsque décision est prise de recourir à des valeurs stockées sous forme d'éléments Wikibase plutôt que sous forme de chaînes de caractères. Ainsi, des données comme les noms des groupes et réseaux de résistance auxquels a appartenu une personne n'ont jamais reposé au CegeSoma sur l'usage d'un vocabulaire contrôlé et étaient simplement encodés tels qu'écrits sur le document ou à l'aide d'abréviations dont l'utilisation n'était ni systématique ni réglementée. Le processus ici consiste donc à partir de l'existant pour créer une liste fermée, agrémentée des formes alternatives de ces noms. Ainsi, plus de 120 noms de groupes ou réseaux de résistance [127] ont par exemple été importés dans l'instance Wikibase afin de pouvoir être utilisés comme valeur de la propriété P74|membre de. Leur existence permet de normaliser et sémantiser l'encodage de cette information. Cependant, il faut garder à l'esprit que cela engendre une charge de travail supplémentaire, dans le sens où les données préexistantes doivent être alignées avec les URIs nouvellement créés et que, idéalement, ces entités de référence présentes dans la Wikibase devraient être davantage décrites, traduites et documentées.

Précisons encore que cet effort de modélisation vise avant tout à constituer un socle de base. En fonction des besoins rencontrés, il pourra ainsi être envisagé de le compléter, par exemple pour inclure des propriétés relatives

123. À savoir : P62|conflit ; P74|membre de ; P76|fonction ; P75|type de résistance ; P77|demande de statut(s) de reconnaissance nationale ; P79|événement-clé ; P78|lieu de détention ; P90|condamné pour ; P92|condamnation.

124. P21|date ; P22|date de début.

125. P29|alias est utilisée ici au sens large pour englober tant des noms de guerre, que des noms de code, pseudonymes ou surnoms ; pour approfondir la réflexion sur les pseudonymes dans l'écosystème Wikibase, voir par exemple : https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Books/2018#Pseudonyms ; https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2019/12#Modelling_a_writer's_pen_name ; https://www.wikidata.org/wiki/Topic:V9a9ogz9e6i380oc.

126. P74|membre de ; P68|distinction reçue.

127. Voir les résultats de cette requête : https://tinyurl.com/y4eotpvh.

à l'armée ou à un grade militaire – qui ne correspondaient pas aux besoins rencontrés en consultant les données et ne représentent pas le cœur de métier de l'institution, qui porte plutôt sur l'histoire sociale –, ou encore des propriétés relatives aux conseils de guerre et conseils militaire dans le cadre de la justice militaire [128].

Enfin, il faut souligner que ces démarches de modélisation posent toutefois la question des données devant être encodées, et de quelle façon : de même que les archivistes sont soumis à un code de déontologie [129], les enjoignant à l'impartialité [130], il semble souhaitable que le personnel chargé de l'implémentation technique de ce type de modélisation nourrisse le même souci de distance et de retenue [131], à défaut de prétendre à une impossible objectivité.

4.3.3 Relations

Cette troisième sous-section englobe toutes les relations pouvant lier une entité à d'autres éléments [132]. Comme le montre la figure 4.7, le concept de *relations* est entendu ici au sens large et inclut des relations de différentes natures : cela peut être la relation à d'autres individus (par exemple dans le cadre de P43|sœur ou frère), la relation à d'autres autorités (par exemple dans le cadre de P47|identifiant ISNI), ou encore le lien à des fichiers (P59|issu de) ou documents d'archives (P72|producteur de).

128. Pour l'instant, un procès peut être décrit à l'aide de la propriété P79|événement-clé qui recevra comme valeur Q6814|procès, et sera éventuellement précisé à l'aide d'un qualificatif de date ou de lieu. Ce système semble suffisant dans un premier temps, sachant que les corpus de données consultés ne contenaient pas de données à ce sujet. Cependant, une modélisation plus fine, tenant compte de la cartographie des conseils de guerre et conseils militaires* pourrait être souhaitable à l'avenir, par exemple dans le cadre du projet BRAIN 2.0 POSTWAREX – dont le CegeSoma est l'un des partenaires – qui veut « étudier de manière approfondie le phénomène de la peine de mort et de son exécution du point de vue de la justice militaire » (CegeSoma, 2020d). *Voir par exemple cette vue récapitulative proposée sur la plateforme Belgium WWII : https://www.belgiumwwii.be/belgique-en-guerre/articles/conseil-de-guerre-repression.html.

129. Publié par l'ICA en 1996 (International Council on Archives, 1996a).

130. Extrait de l'article 3 : « Les archivistes trient les documents avec impartialité, en fondant leur jugement sur une profonde connaissance des exigences administratives et des politiques d'acquisition de leurs institutions. Ils classent et analysent les documents choisis pour être retenus en accord avec les principes archivistiques (en particulier le principe de provenance et le principe du classement d'origine) et les normes universellement reconnues, et ce aussi rapidement que possible. » (International Council on Archives, 1996a, p. 2).

131. Lovink écrivait ainsi, en 2008, en parlant de la hiérarchisation du réel, dans le cadre de la nouvelle génération de moteurs de recherche prenant en charge des requêtes en langage naturel : « however, we may assume that computational linguists will be cautious about acting as a *content police force* that decides what is and what is not crap on the Internet » (Lovink, 2008).

132. Élément est utilisé ici au sens large et non pas au sens strict d'élément Wikibase.

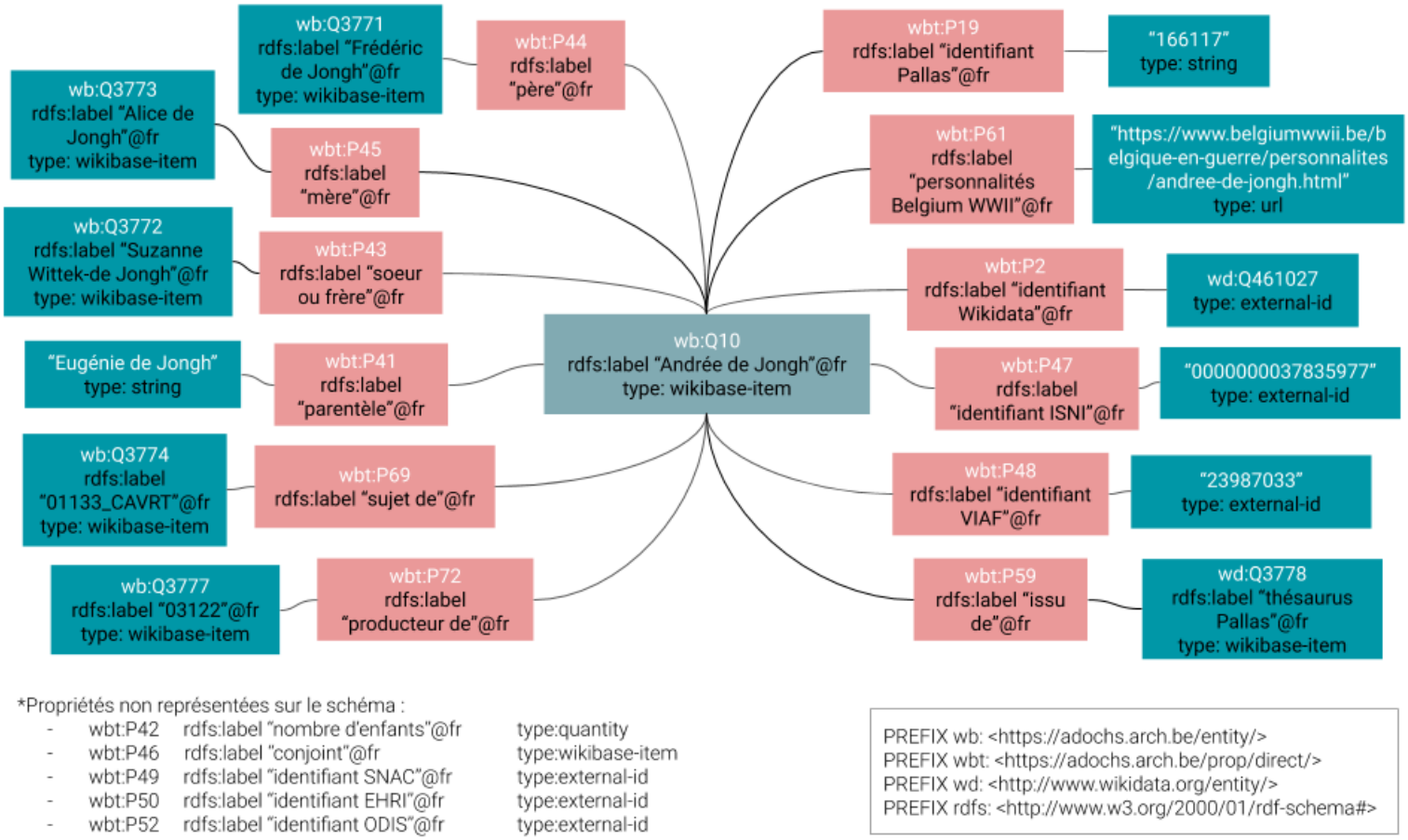


FIGURE 4.7 – Schéma du modèle de données utilisé pour décrire les relations liant une entité Wikibase à d'autres personnes ou ressources ; exemple portant sur la résistante belge Andrée de Jongh|Q10.

À nouveau, les propriétés utilisées se distinguent par le type de données qu'elles acceptent comme valeurs : éléments Wikibase [133], chaîne de caractères [134], quantité [135], URL [136] ou encore identifiants externes [137].

Plusieurs précisions doivent être apportées ici. En ce qui concerne les relations entre personnes, le principe est de ne créer un élément Wikibase pour la personne concernée que si cela ajoute une réelle plus-value : ainsi, la simple mention du nom d'un membre de la famille ne fera pas l'objet d'une nouvelle fiche Wikibase s'il s'agit d'un illustre inconnu, étant donné que cela serait source d'ambiguïté et ouvrirait la porte aux doublons. Dans ce genre de cas où il s'agit d'inconnus ou de liens familiaux flous, il est toutefois possible d'éviter de complètement renoncer à cette information en la documentant à l'aide de la propriété P41|parentèle et d'une valeur composée d'une chaîne de caractères [138]. En revanche, s'il s'agit d'une personne elle-même liée à l'un des grands conflits du XX^{e} siècle et, ou, qu'elle est susceptible d'être liée à d'autres éléments de la Wikibase, et que suffisamment de données d'identification sont connues, une nouvelle entité Wikibase pourra alors être créée pour elle [139]. Il est important de relever que, ce faisant, la base de connaissance va progressivement s'enrichir de données concernant des personnes ne possédant pas forcément de liens directs avec les collections du CegeSoma !

En ce qui concerne les relations à d'autres autorités, deux cas peuvent être distingués : premièrement, il y a les identifiants *internes*, stockés pour des raisons pratiques, afin de maintenir un lien entre les identifiants issus de fichiers préexistants [140] et la nouvelle entité stockée dans la Wikibase ; deuxièmement, il y a les identifiants externes. Ces derniers permettent d'éta-

133. À savoir : P43|sœur ou frère ; P44|père ; P45|mère ; P46|conjoint ; P59|issu de ; P69|sujet de ; P72|producteur de.
134. P19|identifiant Pallas ; P41|parentèle.
135. P42|nombre d'enfants.
136. P61|personnalités Belgium WWII.
137. P2|identifiant Wikidata ; P47|identifiant ISNI ; P48|identifiant ISNI ; P48|identifiant VIAF ; P49|identifiant SNAC ; P50|identifiant EHRI ; P52|identifiant ODIS.
138. C'est le cas par exemple pour la tante d'Andrée de Jongh, Eugénie de Jongh, dont seul le nom est connu. Notons qu'il pourrait par ailleurs être intéressant à l'avenir de compléter cette propriété à l'aide d'un qualificatif permettant de préciser ce lien de parenté.
139. C'est le cas par exemple du père de Andrée de Jongh, Q3771|Frédéric de Jongh (https://adochs.arch.be/wiki/Item:Q3771), qui s'est investi aux côtés de sa fille dans le mouvement de Comète et a été exécuté dans le cadre de ses activités de résistant. En revanche, il faut concéder que le cas de la mère de Andrée de Jongh, Q3773|Alice de Jongh (https://adochs.arch.be/wiki/Item:Q3773), dont l'élément Wikibase a été créé avant tout pour des raisons pragmatiques – à des fins d'illustration de ce modèle de données –, est plus discutable, étant donné l'absence de données d'identification (comme une date de naissance ou de décès), au-delà de ses liens familiaux connus avec Frédéric et Andrée de Jongh.
140. Précisons qu'ils sont destinés à être utilisés de façon complémentaire à la propriété P59|issu de, qui se contente d'établir un lien avec le jeu de données duquel est issue une entité.

blir un lien d'équivalence entre deux entités. C'est avec la création de ces liens que peut pleinement se déployer la dimension de données liées, caractéristique du Web sémantique : il s'agit de faire savoir aux machines que deux URIs se rapportent à une même personne en créant une sorte de raccourci d'une entité vers l'autre. La saisie de cette information passe par l'utilisation d'un identifiant externe dans le cadre du modèle de données Wikibase. Concrètement, la création de ces alignements est loin d'être triviale en matière de temps et de ressources à investir. C'est pourquoi, bien que nous incluions ici différents identifiants afin de prendre en charge plusieurs types de scénarios [141], nous préconisons de concentrer les efforts d'alignement vers Wikidata, dont le rôle de *linking hub* – déjà souligné au cours des chapitres précédents – rend possible la récupération d'une multitude d'identifiants externes à la demande, d'une simple requête SPARQL.

Enfin, l'un des éléments les plus stratégiques dans un contexte archivistique concerne la création et la maintenance de liens entre une autorité et les documents des collections auxquelles elle est liée. Cette problématique est par ailleurs transversale à tout encodage en RDF : il s'agit de déterminer quelle déclaration doit être répétée dans chacune des entités impliquée dans une déclaration, sachant que des requêtes permettent de recréer ces liens à la demande à partir d'une seule déclaration.

Cet exemple issu d'une page de discussion de l'instance FactGrid [142] permet d'imager cela de façon très concrète : en août 2019, des questionnements sont partagés sur la façon la plus pertinente d'organiser l'information : « Repeated information - avoid it or go for it ? » (FactGrid, 2019) Sachant que la base de données contient des milliers de documents, accompagnés de la mention des auteurs, des destinataires, ainsi que des éventuelles personnes mentionnées dans le document, fallait-il reprendre ces informations sur l'élément dédié à une personne ? Par exemple, fallait-il répéter sur la page dédiée à Johann Joachim Christoph Bode, qu'il était l'auteur de 375 documents, premier destinataire de 1 584 documents et mentionné dans 108 autres documents, sous prétexte de présenter à l'utilisateur les données les plus riches possibles [143] ? Avant de conclure en déclarant que la situation idéale serait que chaque fait ne soit présent qu'une seule fois dans la base de

141. Par exemple, pour le cas où une entité serait présente sur la base de données ODIS, mais pas sur Wikidata.

142. La page a été éditée depuis et n'affiche désormais plus ces éléments de réflexion, mais nous pouvons retrouver les informations concernées en tirant profit de l'historique de cette page : https://database.factgrid.de/w/index.php?title=FactGrid:Troubleshooting&diff=1990988&oldid=1662570.

143. « It might be interesting to do this because this is the point at which we show that we have far more information on the man that any other databases. » (FactGrid, 2019).

données [144], et que le reste soit affiché à la demande, à l'aide de requêtes ou d'interfaces permettant par exemple de voir tous les documents dont Bode est l'auteur.

Dans le cas relatif à FactGrid, les documents sont donc les éléments centraux. Dans le cas des données du CegeSoma, de telles questions se posent également, vu que les données d'autorité sur les personnes et les données de description des documents d'archives ne sont pas stockées au sein de la même infrastructure. Avant de rentrer dans les détails, nous devons commencer par souligner le caractère extrêmement provisoire des mesures implémentées : étant donné le contexte de transition dans lequel est actuellement plongé le CegeSoma [145], il s'avère que les collections du Centre, destinées à être désormais consultées via le moteur de recherche des Archives de l'État, ne sont pas encore disponibles en ligne par ce biais. Nous présentons donc ici à la fois la solution *ad hoc* adoptée dans le cadre de cette démonstration et la solution qui serait envisagée dans le cadre d'une pérennisation de cette base de connaissance, au terme du processus complet de migration des collections du CegeSoma vers les systèmes de gestion des Archives de l'État.

La solution provisoire, adoptée dans le cadre de cette étude de cas, vise à illustrer de quelle manière les entités Personne présentes dans une instance Wikibase devraient permettre d'accéder à une vue agrégée de toutes les ressources auxquelles elles sont liées. Dans le cadre des collections du CegeSoma, ce lien apparaît sous deux formes distinctes : d'une part, un individu peut avoir été identifié comme le producteur d'un fonds, d'autre part, il peut avoir été lié à l'un ou l'autre niveau d'arborescence dans le cadre de l'indexation d'un fonds d'archives. Ces deux cas de figure sont pris en charge par deux propriétés distinctes, respectivement P72|producteur de [146] et P69|sujet de [147]. Ces propriétés impliquent d'utiliser comme valeur un élément présent dans la Wikibase. Cela signifie que, faute d'inventaires accessibles en ligne et pouvant être désignés à l'aide d'un permalien, nous avons dès lors dû commencer par créer près de 3 000 éléments Wikibase [148]

144. *We should otherwise not be to [sic] eager to begin this because this was the reason why we left the Wiki and went for the database : We do not want to constantly mirror information in various data sets. We do not want the user to keep in mind on which pages a certain fact has to be mirrored.* (FactGrid, 2019).
145. Voir Chapitre 3.
146. https://adochs.arch.be/wiki/Property:P72.
147. https://adochs.arch.be/wiki/Property:P69.
148. Les utiliser sous forme d'éléments plutôt que de chaînes de caractères permet de faciliter la saisie de ces informations à l'avenir, d'inclure tant leur cote que leur titre à la fois en français et en néerlandais, de faciliter la formulation de requêtes associées aux fonds et enfin d'éventuellement inclure des informations additionnelles comme une URL.

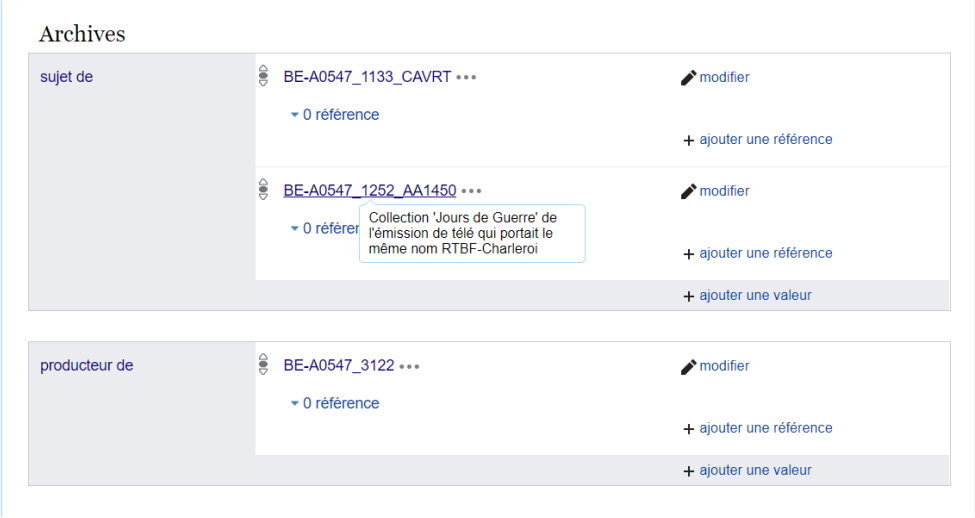


FIGURE 4.8 – Exemple d'archives liées à la résistante belge Andrée de Jongh|Q10 par le biais des propriétés P69|sujet de et P72|producteur de.

basés sur les identifiants uniques associés aux *blocs d'archives* [149] correspondants dans SAM, le système de gestion des Archives de l'État. À ces éléments ont été associés les noms des blocs en français ou en néerlandais [150]. Cela signifie qu'une personne effectuant des recherches sur la co-fondatrice du réseau Comète – Andrée de Jongh –, pourra voir d'un seul coup d'œil les archives auxquelles elle est associée, comme le montre la figure 4.8.

Il faut noter que ce système – tout à fait transitoire, nous le rappelons – est toutefois loin d'être optimal, dans la mesure où il ne tient pas compte du niveau de granularité auquel l'indexation a été réalisée [151], mais surtout dans la mesure où il est très exigeant au niveau de la maintenance des données. En effet, il implique non seulement de multiplier des informations identifiques – les cotes de références –, qui devront dès lors être tenues à jour au sein de systèmes différents – Wikibase et SAM –, mais également de devoir créer des liens entre documents et autorités à deux reprises : une première fois de l'inventaire vers l'URI Wikibase associée à la personne ; une seconde fois de la fiche Wikibase vers l'inventaire correspondant. Un tel dis-

149. Cette terminologie utilisée par les Archives de l'État désigne « un ensemble d'archives constituant l'unité matérielle de base pour la gestion d'un dépôt d'archives (tant analogique que numérique) » (Lardinois *et al.*, 2019, p. 9).

150. La liste des 2 827 blocs d'archives peut être consultée à l'aide de cette requête : https://tinyurl.com/y2o6ccg8.

151. Cela fait malheureusement partie des dommages occasionnés par la migration des données vers les Archives de l'État : il n'existe pas – pour l'heure, du moins – de table de correspondance entre les niveaux de description les plus détaillés de Pallas et les nouvelles cotes ayant été attribuées dans SAM, qui se limitent pour l'instant à un niveau plus générique de description.

positif génère une redondance inutile de l'information et accroît considérablement la quantité de travail d'encodage et de maintenances des données, tout en augmentant le risque d'erreurs.

Pour toutes ces raisons, le scénario envisagé en cas de pérennisation de cette instance Wikibase pour la gestion des autorités consisterait à indexer les inventaires d'archives [152] au niveau de granularité approprié en intégrant dans l'EAD [153] les URIs – des personnes concernées – issus de la Wikibase. Ces inventaires EAD étant indexés par Solr, le logiciel derrière le moteur des recherches des Archives de l'État, il serait ensuite possible de générer une page dynamique listant tous les titres et niveaux de description contenant l'URI associé à une certaine personne [154], de même que les éventuels documents numérisés [155]. La base de connaissance servirait donc à la fois à identifier une personne de façon non ambiguë et à entreposer des données structurées multilingues pouvant être récupérées pour être affichées sur la page du moteur de recherche des Archives dédiée à une certaine personne. La base de connaissance pourrait dès lors inclure un permalien renvoyant vers cette page, mais elle n'aurait pas pour mission de constituer le lieu de consultation des données relatives aux documents d'archives.

Enfin, notons qu'au-delà de la distinction déjà opérée entre *producteur* ou *sujet* de documents d'archives, il pourrait par ailleurs être intéressant de prendre en charge un niveau de granularité plus fin (comme par exemple en distinguant *creator*, *collector* et *compiler* [156]).

4.3.4 Alignements avec Wikidata et RiC-O

Comme évoqué dans l'introduction de cette section, l'objectif était de pouvoir réutiliser au maximum des propriétés Wikidata préexistantes, afin de favoriser l'interopérabilité des données, notamment dans la perspective d'un écosystème Wikibase permettant des échanges et des réutilisations de données entre différentes instances. Ainsi, comme le montre de façon synthé-

152. À l'aide de la balise EAD <persname>.

153. Il s'agit ici d'adopter une vision à relativement court terme ; à plus long terme, les Archives de l'État sont susceptibles d'adapter leurs pratiques de description à la nouvelle norme Records in Contexts, dans ce cas, cette étape d'indexation ne prendra peut-être plus place dans un inventaire EAD.

154. À l'instar de ce que proposent par exemple les Archives nationales de France, voir : https://francearchives.fr/fr/agent/217549846.

155. En particulier les nombreuses collections photographiques du CegeSoma.

156. Nous pouvons également aux exemples proposés par Clavaud qui explique, dans le contexte des systèmes d'informations des Archives nationales de France, que devraient être pris en compte non seulement les producteurs initiaux de documents, mais également les personnes physiques ou morales les ayant produits, accumulés, maintenus, ainsi que les auteurs ou responsables intellectuels, les expéditeurs et destinataires initiaux, ainsi que les personnes qui en sont le sujet (Clavaud, 2019b).

FIGURE 4.9 – Illustration de la proportion de propriétés présentes dans la Wikibase ayant pu être alignées avec les propriétés Wikidata ou RiC-O.

tique [157] la figure 4.9, la plupart des propriétés présentes dans la Wikibase possèdent leur équivalent Wikidata [158]. Cette figure montre également les efforts déployés pour aligner les propriétés de la Wikibase avec l'ontologie Records in Contexts (RiC-O) [159].

Concrètement, sur un total de 72 propriétés actuellement [160] utilisées dans la Wikibase, 55 d'entre elles ont pu être alignées vers une propriété Wikidata équivalente, soit 76%, et 52 ont été alignées vers une propriété RiC-O équivalente, soit 72%. Ces pourcentages sont toutefois à considérer avec précaution, dans la mesure où toutes les propriétés ne se superposent

157. Pour une vue plus détaillée, consulter le tableau présenté dans l'Annexe 6, page 357, qui reprend chacune des propriétés avec son libellé et ses liens d'équivalence.
158. Ces liens ont été documentés sur la Wikibase à l'aide de la propriété P14|propriété équivalente, qui correspond elle-même à une propriété équivalente à *owl :equivalentProperty*.
159. Nous tenons à remercier chaleureusement Florence Clavaud, membre exécutif du groupe ICA-EGAD et responsable de l'équipe de développement de RiC-O, pour ses divers éclairages et contributions, notamment sous la forme de commentaires sur notre tentative d'alignement entre nos propriétés locales et les propriétés RiC-O. Ces commentaires – qui sont venus confronter ou compléter nos propres observations – sont repris dans le tableau proposé dans l'Annexe 6, page 357.
160. En août 2020.

pas avec la même exactitude et qu'une certaine tolérance a été adoptée. Comme nous le verrons à l'aide d'exemples au cours des prochains paragraphes, le degré d'équivalence varie énormément. En effet, il faut garder à l'esprit que les propriétés utilisées dans le cadre de notre instance Wikibase ou dans le cadre de Wikidata sont destinées à répondre à des usages spécifiques, là où l'ontologie RiC-O est une ontologie de haut niveau qui se doit d'être assez générale pur pouvoir être réutilisée à travers différentes mises en œuvre. Cela signifie que dans un certain nombre de cas, les propriétés de la Wikibase représentent en réalité des sous-propriétés en puissance de propriétés RiC-O.

Les propriétés n'ayant pu être alignées avec Wikidata peuvent être regroupées en deux principales catégories : d'une part, les propriétés à usage interne [161], d'autre part, des propriétés plus spécifiques au contexte archivistique [162] ou historique [163]. En ce qui concerne les cas d'équivalences imparfaites, nous avons fait preuve d'une certaine tolérance à quelques reprises, lorsque les propriétés Wikibase sont une réplique des propriétés Wikidata, se distinguant uniquement par leur *data-type* leur a été attribué [164].

En ce qui concerne l'ontologie Records in Contexts (RiC-O) [165], il ne s'agit pas non plus toujours d'équivalences parfaites. En effet, l'alignement d'ontologies est connu pour être une tâche ardue, en raison du fait que les schémas des bases de connaissance peuvent varier de façon significative « in terms of subject area coverage, level of abstraction, ontology modeling philosophy, and language » (Zhou *et al.*, 2020, p. 1). Cette variation dans le degré d'abstraction signifie que nous avons fait preuve d'une certaine flexibilité lors

161. Qu'il s'agisse de propriétés directement liées au fonctionnement de l'instance Wikibase, comme par exemple P3|élément(s) à combiner, ou d'identifiants internes, propres au CegeSoma, comme P19|identifiant Pallas.

162. Par exemple, la propriété Wikibase P72|producteur de, n'existe pas à part entière sur Wikidata, étant donné que le lien est documenté sur Wikidata dans le sens inverse : ce sont les documents d'archives qui ont pour P6241|producteur une personne physique ou morale.

163. Ainsi, le fait que Wikidata soit une base de connaissance à vocation généraliste implique par exemple que les résistants y soient simplement signalés à l'aide de la propriété P106|occupation – également utilisée pour les professions – accompagnée de la valeur Q23833535|résistant. C'est différent dans le cadre des données du CegeSoma. En effet, étant donné qu'il s'agit du cœur de métier du Centre, il s'agit d'indiquer cette information de façon plus fine, à l'aide de données factuelles connues : par exemple la propriété P75|type de résistance permet d'indiquer à quelle forme d'activité de résistance s'est livrée une personne, tandis que la propriété P77|demande de statut(s) de reconnaissance nationale, permet de documenter le fait qu'une demande de reconnaissance de statut de résistant a été adressée auprès des autorités belges.

164. C'est le cas par exemple de la propriété P85|affirmé (https://adochs.arch.be/wiki/Property:P85) qui a comme *data-type* des chaînes de caractères, là où la propriété Wikidata originale accepte comme valeurs des éléments.

165. Nous nous référons ici à la version v0.1, publiée par l'International Council on Archives en décembre 2019 (International Council on Archives Expert Group on Archival Description, 2019a), qui constitue la représentation formelle de la version v0.2 du modèle conceptuel Records in Contexts (RiC-CM).

de l'établissement des relations d'équivalence. Ainsi, lorsqu'il s'avère qu'une propriété Wikibase pourrait potentiellement être *traduite* à l'aide d'une sous-propriété – en puissance – de RiC-O, nous avons considéré qu'un lien d'équivalence pouvait être instauré entre la propriété Wikibase plus spécifique et la propriété RiC-O plus générique [166]. À quelques reprises, c'est le cas de figure inverse qui s'est présenté [167]. Enfin, les propriétés OWL ou RDFS dont RiC-O se sert, – comme *rdfs :seeAlso* – ont été inclues dans cette tentative d'alignement, bien qu'elles ne soient pas propres à RiC-O.

Pour conclure, précisons que si ce travail d'alignement a été réalisé ici manuellement à des fins d'illustration et de réflexion sur la complémentarité et la spécificité de chacune des ontologies, il pourrait toutefois être intéressant de poursuivre ce processus de manière plus aboutie. Que ce soit en travaillant sur la formalisation de cet alignement à l'aide du format EDOAL (Expressive and Declarative Ontology Alignment Language) [168], à l'instar de Zhou *et al.*, qui ont aligné les éléments et propriétés de l'instance Wikibase Enslaved [169] avec les classes et propriétés OWL de l'ontologie Enslaved [170] (Zhou *et al.*, 2020), ou encore en passant par des algorithmes d'alignements complexes tels que testés par exemple dans le cadre de la Conférence Ontology Alignment Evaluation Initiative [171]. Il faut toutefois noter qu'une telle démarche pourrait être freinée par le fait qu'il n'est pour l'instant [172] pas possible de récupérer le modèle interne d'une certaine instance Wikibase en OWL/RDF [173].

166. Cela explique par exemple pourquoi l'ensemble des identifiants stockés à l'aide d'une propriété Wikibase sont indiqués comme possédant un équivalent RiC-O : bien qu'ils n'existent pas en tant que tels, ils pourraient aisément être englobés dans des sous-propriétés de *rico :identifier*. Pour d'autres exemples, voir l'Annexe 6, page 357.
167. Ainsi, la définition de *rico :certainty* s'avère par exemple plus restreinte que notre propriété P25|qualité de l'information.
168. http://alignapi.gforge.inria.fr/edoal.html.
169. https://lod.enslaved.org/.
170. https://docs.enslaved.org/ontology/.
171. Voir par exemple les résultats pour l'édition 2019 : Algergawy *et al.* (2019).
172. La communauté Wikidata nourrit une réflexion à ce sujet par le biais d'un WikiProject Ontology/Modelling, voir : https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Modelling.
173. Comme nous l'a fait remarquer Florence Clavaud via une communication personnelle.

5 | Implémentation d'une instance Wikibase

Introduction

Ce chapitre s'intéresse à l'implémentation de notre instance Wikibase. En effet, maintenant que les données ont été préparées et qu'un modèle a été conçu afin d'englober les différents cas de figure pouvant se présenter, les notices d'autorité dédiées aux personnes vont pouvoir être créées directement dans la Wikibase. Elles seront ainsi toutes intégrées dans une seule et même infrastructure de données et chaque entité sera désignée à l'aide d'un identifiant unique et persistant. Ces données d'autorité devraient pouvoir être plus facilement gérées, interrogées, partagées et réutilisées : c'est ce que nous allons tester au cours de ce cinquième chapitre.

L'état de l'art l'a montré : depuis que la diffusion du logiciel Wikibase est considérée comme l'une de priorités de l'équipe de développement de Wikidata, de multiples efforts sont déployés pour faciliter son installation et répondre aux difficultés rencontrées par les utilisateurs. Ces efforts se traduisent notamment par la mise en ligne de nouvelles versions de l'image Docker permettant d'installer Wikibase. Étant donné ces fréquents changements, cette thèse ne s'apesantit pas sur le processus technique qui a permis de créer le prototype au cœur de cette étude de cas [1]. Dans ce chapitre, l'attention est davantage portée sur le workflow [2] pouvant être mis en place pour gérer les données d'autorité, ainsi que sur l'exploration de différents cas d'usage. La première section, dédiée au workflow de gestion des données, est structurée autour de deux étapes-clés : la création et l'administration des données. La seconde section vise quant à elle à examiner quatre cas

1. Des informations complémentaires sont toutefois fournies en annexes : l'Annexe 7, page 357, décrit l'installation de notre instance Wikibase sur les serveurs des Archives de l'État, tandis que l'Annexe 8, page 367, présente les détails de configuration de cette instance.

2. Traduit littéralement, ce terme emprunté à l'anglais désigne un *flux de travaux*. Étant donné que cet anglicisme est entré dans l'usage, nous l'utilisons ici en anglais et sans emphase, afin de favoriser la lisibilité du texte.

d'usage : la recherche et l'accès aux données ; la réutilisation des données ; l'exploration du potentiel des requêtes SPARQL fédérées pour compléter les informations contenues dans l'instance Wikibase en les enrichissant de données externes ; la réconciliation de noms de personnes à partir des données contenues dans la Wikibase.

5.1 Workflow

5.1.1 Création des données

Comme le montre la figure 5.1, la première étape, dédiée à la création des données recèle à elle seule une certaine complexité. Ce processus peut en effet être décomposé en trois temps distincts que nous allons tenter de décrire et d'illustrer à l'aide d'exemples au cours des prochains paragraphes. Le premier temps correspond à la création des entités Wikibase permettant de mettre en place le modèle de données élaboré au cours des pages précédentes. Cela inclut les propriétés en tant que telles, mais également les données de base [3] nécessaires au fonctionnement de ce modèle. Le deuxième temps correspond au chargement initial des jeux de données préexistants, que nous distinguons du troisième temps, à savoir l'ajout continu de nouvelles données dans l'instance Wikibase.

En ce qui concerne la première phase de création des données, elle vise à doter la base de connaissance des éléments nécessaires à la description des données. Elle repose sur le modèle de données détaillé au cours des pages précédentes. Cette étape va donc de la création de propriétés [4] telles que P31|date de naissance [5], à l'ajout des éléments de base permettant par exemple de spécifier la P1|nature d'un élément [6] à l'aide d'une valeur comme Q3617|personne [7], en passant par les éventuels compléments nécessaires pour affiner ou nuancer un triplet, tel qu'un qualificateur décrivant la source de l'information [8].

La figure 5.2 montre un exemple des propriétés et éléments nécessitant d'avoir été préalablement créés pour que la notice d'autorité Wikibase sur la résistante belge Andrée de Jongh puisse être complétée. Comme le montre cet exemple et comme nous l'avons vu au cours de la section précédente dédiée à la modélisation des données, c'est surtout le type de données [9]

3. En anglais, nous avons rencontré le terme *core items* (Miller, 2018).
4. Plus précisément la création d'un libellé, d'une description et d'éventuels alias dans une ou plusieurs langues, potentiellement accompagnés de déclarations.
5. https://adochs.arch.be/wiki/Property:P31.
6. https://adochs.arch.be/wiki/Property:P1.
7. https://adochs.arch.be/wiki/Item:Q3617.
8. Comme P60|importé de Wikidata : https://adochs.arch.be/wiki/Property:P60.
9. Voir : https://www.wikidata.org/wiki/Help:Data_type/fr.

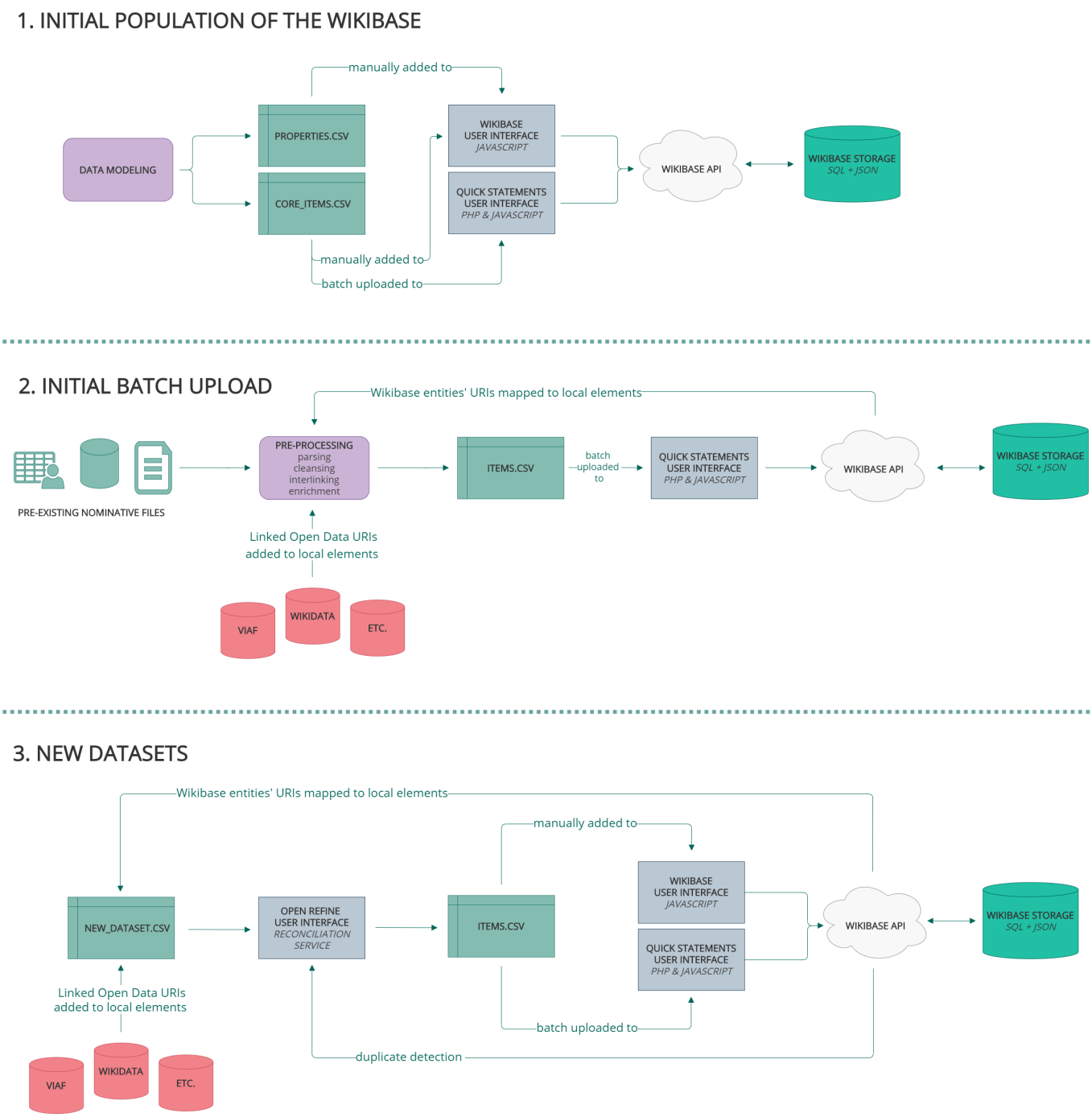


FIGURE 5.1 – Les trois temps du chargement de données dans l'instance Wikibase.

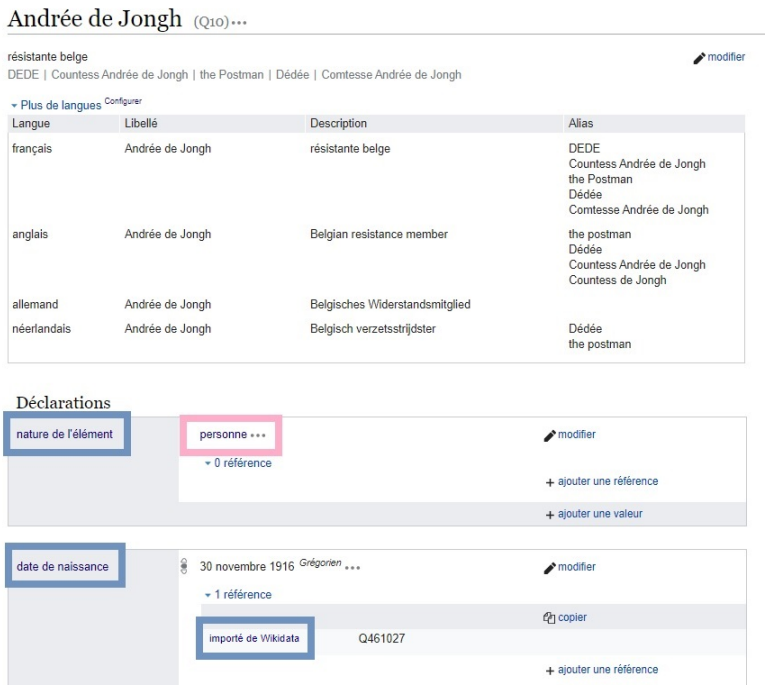


FIGURE 5.2 – Avant de pouvoir encoder les informations décrivant l'élément Q10|Andrée de Jongh, il a fallu créer des propriétés (en bleu) telles que *nature de l'élément*, *date de naissance* ou *importé de*, ainsi que l'élément (en rose) *personne*, permettant de spécifier sa nature.

associé à une propriété qui est déterminant. Ainsi, la propriété P31|date de naissance requiert comme type de données un *point dans le temps* : nul besoin donc d'importer en amont dans la base de connaissance toutes les dates possibles, il suffira d'encoder ou d'importer une date dans un format adéquat au moment de la création des données [10]. Par ailleurs, il faut noter qu'il n'existe pas de franche délimitation entre les données *de base* et celles qui composent les jeux de données de l'institution. Par exemple, les noms de lieux constituent à la fois des données d'autorité susceptibles d'être publiées et mises à disposition via la Wikibase [11], et des éléments de base utiles pour décrire d'autres données, comme le lieu de naissance d'un individu.

Dans le cadre du lancement d'une nouvelle instance Wikibase, il est possible d'opter pour une importation de données directement issues de Wiki-

10. Par contraste, pour une propriété comme P38|pays de nationalité, davantage de choix de types de données se présentent potentiellement : du texte – sous forme de *chaîne de caractères* ou de *texte monolingue*, selon que la valeur doive être traduite ou non –, une URL, un identifiant externe – comme un élément Wikidata – ou encore un élément de la Wikibase elle-même – comme par exemple Q31|Belgique. Si c'est ce dernier type qui est choisi, les éléments devront être créés en amont.

11. À l'instar de Biblissima qui a choisi Wikibase pour partager son *référentiel des lieux géographiques* : https://data.biblissima.fr/w/Référentiel_des_noms_géographiques.

data. La célèbre base de connaissance est en effet équipée d'une ontologie et de données libres de droit, ayant vocation à décrire rien de moins que *le monde* [12]. Ce caractère généraliste encourage la réutilisation de ces données : plutôt que de tout recréer de zéro en prenant en charge la maintenance de données *annexes* ne faisant pas l'objet de son cœur de métier, une institution peut faire le choix de réutiliser ce qui existe déjà ailleurs. Comme nous l'avons vu au cours du chapitre précédent, cette pratique favorisant l'interopérabilité est particulièrement encouragée dans le cadre du choix de propriétés et nous a ainsi poussée à élaborer notre modèle de données en reprenant des propriétés Wikidata existantes dès que c'était possible.

À l'heure actuelle, les possibilités concrètes de partage et de synchronisation de données entre instances Wikibase s'avèrent toutefois encore limitées. Une option encouragée au sein des premiers tutoriels dédiés à Wikibase (Scott et Allison-Cassin, 2018) consiste donc à utiliser un processus automatisé permettant l'import des entités Wikidata de son choix. Cette étape peut par exemple être réalisée à l'aide d'une extension MediaWiki appelée WikibaseImport [13], qui est équipée d'un script permettant d'importer automatiquement tout ou partie des entités d'une autre instance Wikibase, qu'il s'agisse de Wikidata ou d'une autre instance.

Cependant, si réutiliser les données Wikidata peut aparaître comme un gain de temps à première vue, la question de la maintenance ne doit une fois de plus pas être négligée. Premièrement, elle se pose d'abord au niveau de l'outil d'import lui-même. Ainsi, il apparaît que le script susmentionné ne fonctionne pas correctement, du moins lorsqu'il est utilisé dans le cadre d'instances Wikibase-Docker récentes [14]. Deuxièmement, la question de la maintenance surgit au niveau des données mêmes, étant donné que les données contenues dans Wikidata ne cessent d'évoluer au gré de l'activité de la communauté (Allison-Cassin et Seeman, 2019). Imaginons par exemple l'importation de tous les pays répertoriés par Wikidata, accompagnés de déclarations décrivant leurs spécificités. La liste de chefs d'État [15] de l'élément Q31|Belgique [16] – dont la figure 5.3 montre un extrait – serait par exemple importée telle quelle. Par défaut [17], rien n'assurera l'actualisation de ces don-

12. https://be.wikimedia.org/wiki/Wikidata.
13. https://github.com/filbertkm/WikibaseImport.
14. Ainsi, bien qu'ayant utilisé ce dernier avec succès au cours de l'hiver 2018, nous avons constaté, lors d'une utilisation dans le cadre d'une nouvelle instance créée en janvier 2020, que toutes les données n'étaient pas importées correctement (voir : https://github.com/filbertkm/WikibaseImport/issues/19#issuecomment-62519929 et https://phabricator.wikimedia.org/T209803).
15. P35|chef d'État : https://www.wikidata.org/wiki/Property:P35.
16. https://www.wikidata.org/wiki/Q31.
17. Il faut savoir que l'extension WikibaseImport prévoit une table de correspondance entre anciens et nouveaux identifiants numériques, voir : https://github.com/filbertkm/WikibaseI

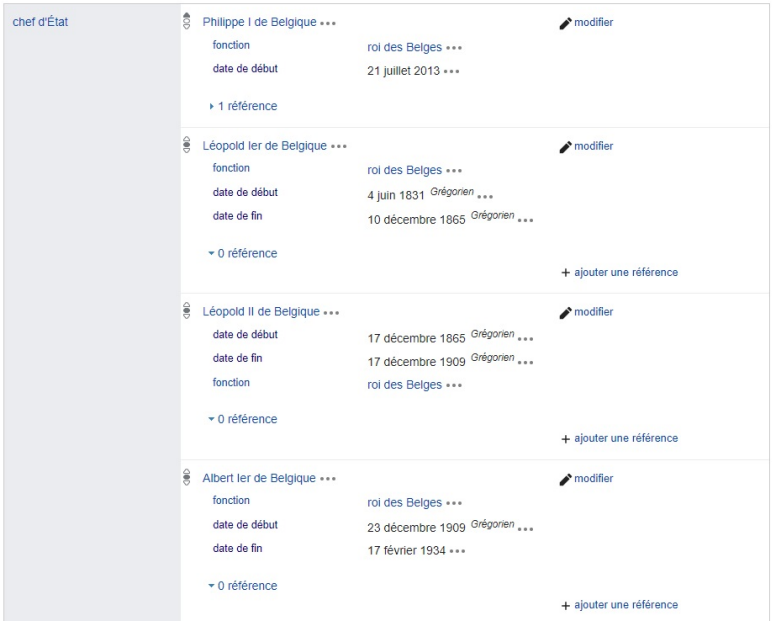


FIGURE 5.3 – Extrait des chefs d'État documentés sur la fiche Wikidata de l'élément Q31|Belgique. Source : Wikidata (https://www.wikidata.org/wiki/Q31).

nées le jour où un nouveau chef d'État devra être désigné. Il en va de même si certaines données sont rectifiées, complétées ou supprimées sur Wikidata.

Pour résoudre la question de la synchronisation entre les données de deux instances Wikibase, il est certes possible de façonner un système de mise à jour quotidienne des données de la Wikibase afin de bénéficier des éventuelles modifications ayant été effectuées sur Wikidata, à l'instar de ce qui a été implémenté dans le cadre de l'instance *the EU Knowledge Graph* (De Wilde, 2020). Cela suppose toutefois de réelles compétences techniques et une certaine disponibilité pour assurer le suivi de telles opérations. De plus, cela nécessiterait d'approfondir ce qu'il se passe dans des situations particulières, comme lors de la suppression d'éléments Wikidata qui auraient été utilisés localement comme *valeurs* de nouvelles déclarations.

Enfin, outre ces réticences liées à des aspects d'ordre technique, le processus peut également être questionné en raison de ses implications au niveau du contenu lui-même. La question est de savoir si cela est souhaitable de *tout* importer en masse, sans filtre, sachant que, d'une part, cela entraîne une évidente perte de contrôle en ce qui concerne la qualité des données, déjà évoquée au cours des pages précédentes, et que, d'autre part, cela peut

mport/blob/master/README.md, un système de synchronisation peut donc potentiellement être implémenté.

également générer un bruit non souhaitable [18]. Pour toutes ces raisons, nous avons dès lors renoncé à une récupération automatisée de données issues de Wikidata.

Pour ajouter de nouvelles données dans la Wikibase, deux grandes options se présentent. Premièrement, une interface graphique permet de créer manuellement une nouvelle entité, en commençant par encoder son label et sa description – dans une seule langue pour commencer –, qui peuvent potentiellement être complétées à l'aide de formes alternatives du nom – par le biais d'*alias* –, mais également par des déclarations venant décrire cette entité. Chacune de ses déclarations peut en outre être détaillée à l'aide de qualificateurs ou de références. Si Wikibase offre la possibilité à toute personne disposant des droits correspondants de créer et modifier les données (ouvertes et liées) de la Wikibase en temps réel sans disposer de compétences techniques particulières, il faut toutefois relever une certaine *régression*, dans la mesure où le logiciel n'offre pour l'instant pas la possibilité de déployer des listes autodéroulantes restreignant les choix possibles lors de l'encodage d'une valeur, contrairement aux logiciels traditionnels de description de collections permettant d'associer à certains champs des listes de termes contrôlés. Par ailleurs, l'interface Wikibase ne permet pas non plus d'afficher de formulaire permettant de restreindre les informations devant être remplies pour les éléments d'une certaine nature (comme par exemple les personnes) [19].

Plus généralement, bien que l'interface soit équipée d'une fonctionnalité d'autocomplétion permettant de fluidifier ce processus, il est clair qu'ajouter manuellement chaque information l'une après l'autre n'est pas adapté à un import massif de nouvelles données, pouvant dans certains cas se compter en dizaines de millions [20]. Pour de telles importations, il est donc possible de préparer un jeu de données qui pourra être importé par un robot com-

18. Comme le relatent par exemple Allison-Cassin et Seeman : « Property and item import from Wikidata meant many associated properties, items, labels, and dependencies were also added and I ended up deleting most of them. An example from our use case is when I tried to import the province of Ontario as an item into our Wikibase instance and ended up with a property of *motto text* – so there are probably ways to get around this but it's meant to illustrate that it's VERY easy to lose control » (Allison-Cassin et Seeman, 2019).

19. La communauté Wikidata travaille toutefois sur la question, comme en témoigne le développement de l'outil expérimental Cradle, qui vise précisément à pouvoir créer un élément Wikidata à l'aide d'un formulaire, voir : https://cradle.toolforge.org.

20. Ainsi, les collaborateurs de la BNF testant Wikibase doivent par exemple prendre en charge les 60 millions d'éléments du Fichier National d'Entités, voir : https://github.com/abes-esr/poc-fne/issues/18#issuecomment-524294307.

muniquant avec l'API *(Application Programming Interface)* [21] associée à la Wikibase [22].

Comme à d'autres occasions, la stratégie utilisée par la communauté Wikibase consiste à réutiliser des outils ou programmes développés dans le cadre de Wikidata, à l'instar de l'outil QuickStatements [23], des bibliothèques Python Wikidata Integrator [24] ou Pywikibot [25], ou encore de l'outil Javascript Wikibase-Edit [26]. Ces derniers fournissent des scripts permettant à un robot – c'est-à-dire un compte utilisateur doté de droits d'édition spécifiques – de créer tous les labels, descriptions, alias et déclarations nécessaires.

Très concrètement, dans le cadre de notre Wikibase, les propriétés ont été créées manuellement [27] dans l'interface utilisateur Wikibase, tandis que les éléments de base ont été soit ajoutés manuellement, soit importés massivement à l'aide de l'outil QuickStatements [28], comme le montre la figure 5.1.

Une fois les propriétés et les données de base créées, est venu le second temps de la création des données, qui concerne le chargement initial, c'est-à-dire l'importation dans la Wikibase des jeux de données préexistants. Concrètement, les échantillons de données préparés au cours des étapes précédentes ont été adaptés pour tenir compte des identifiants Wikibase des propriétés et éléments récemment créés [29] avant d'être importés [30] dans Wi-

21. « A set of protocol constructs offered by a Web application through which third-party Web or software applications can interact with it » (Verborgh *et al.*, 2015).

22. Dans le cadre de la Wikibase DataCegeSoma, l'API est disponible à cette URL https://adochs.arch.be/w/api.php.

23. https://www.wikidata.org/wiki/Help:QuickStatements/fr.

24. https://github.com/SuLab/WikidataIntegrator.

25. https://www.mediawiki.org/wiki/Manual:Pywikibot.

26. https://github.com/maxlath/wikibase-edit.

27. En effet, sachant que QuickStatements ne prend pas en charge la création de propriétés, et au vu du nombre limité de propriétés devant être créées, des difficultés rencontrées – par exemple pour prendre en charge simultanément différentes traductions et alias – avec les autres outils d'automatisation ayant été passés en revue, nous avons finalement opté pour un encodage manuel.

28. Cet outil est inclu par défaut dans la version *dockerisée* de Wikibase.

29. Par exemple, une colonne de fichier intitulée *pays de nationalité* devient P38. Ces étapes réalisées ici de manière relativement *artisanale*, gagneraient certainement à être systématisées et automatisées.

30. Nous précisons ici que, pour l'instant, seule une minorité des entités issues du premier jeu de données de notre échantillon – à savoir le thésaurus Pallas – ont d'ores et déjà été importées dans l'instance Wikibase, à savoir les entités ayant pu être liées à des entités issues des deux autres jeux de données de notre échantillon. Le reste des données fera l'objet d'une importation massive après la clôture de la vérification des alignements vers Wikidata, un travail destiné à prendre place au cours des mois à venir, comme souligné au cours de la section 4.2.3.

kibase à l'aide de l'interface QuickStatements[31]. Au total, 75 propriétés ont été ajoutées[32], ainsi que 7 087 éléments[33].

Les principales difficultés rencontrées lors de cette étape concernent l'adaptation des données à la syntaxe attendue par QuickStatements[34] et le fait que l'outil ne prenne pas en charge des données contenant des valeurs manquantes, rendant le chargement des lots de données extrêmement laborieux[35].

Enfin, le troisième temps concerne l'ajout continu de nouvelles données dans l'instance Wikibase. Le processus est similaire, si ce n'est que le processus devra inclure la détection de doublons, qui peut par exemple être réalisée par le biais d'un service de réconciliation reposant sur l'utilisation combinée de l'API associée à la Wikibase et du logiciel OpenRefine, comme nous l'illusterons au cours de la section 5.2.4. À nouveau, les données pourront ensuite être ajoutées à la fois sous forme de jeux de données ou manuellement.

5.1.2 Administration des données

Parmi les synonymes du verbe *administrer* figurent[36] les verbes *gérer ; superviser ; coordonner ; régenter ; appliquer ; contrôler*. Cela exprime bien les différentes actions pouvant prendre place au cours de cette deuxième étape du workflow de gestion des données. Une fois les données créées, éven-

31. https://quickstatements-adochs.arch.be/.

32. La liste de ces 75 propriétés, accompagnées de leur description, peut être obtenue à l'aide de cette requête : https://tinyurl.com/y697u3xy.

33. Ces 7 087 éléments sont de 23 *natures* différentes (pour des raisons de clarté, cette liste reprend uniquement les éléments possédant une déclaration basée sur la P1|nature de l'élément, elle ne prend donc pas en compte quelques exceptions, à savoir une minorité d'éléments – a priori, moins de 50 – n'ayant pas encore fait l'objet d'une telle déclaration) : la Wikibase compte ainsi 2 828 cotes de références archivistique ; 2 493 sections de commune situées en Belgique ; 588 communes situées en Belgique ; 472 personnes et 302 localités situées en Belgique, pour ne citer que les plus nombreuses. La liste récapitulative de ces éléments peut être obtenue à l'aide de la requête suivante : https://tinyurl.com/y5o5zuo6, tandis que l'ensemble de ces éléments est visible en suivant cette requête : https://tinyurl.com/y3pmqlcb.

34. Qui est par exemple particulièrement laborieuse dans le cadre de valeurs présentes sous forme de texte littéral, devant être encadrés de double ou triples guillemets avant leur importation, voir : https://www.wikidata.org/wiki/Help:QuickStatements/fr#Syntaxe_de_fichier_CSV.

35. En effet, certains champs de nos jeux de données ne contenaient pas systématiquement de valeurs, ce qui nous a donc amené à également utiliser un autre outil d'importation massive, à savoir le programme csv2wikibase – reposant sur Wikibase-Edit et généreusement mis à notre disposition par Sébastien Beyou, fondateur de Wiki Valley –, qui permettait de pallier ce type de difficultés, mais en a cependant généré d'autres (que nous ne développons pas ici dans un souci de concision), illustrant la difficulté de trouver un seul outil permettant de couvrir l'intégratité de nos besoins.

36. Nous nous référons à la liste des synonymes proposés par le Dictionnaire Électronique des Synonymes que publie le Laboratoire CRISCO (Centre de Recherche Inter-langues sur la Signification des en COntexte) : https://crisco2.unicaen.fr/des/synonymes/administrer.

tuellement modifiées ou complétées, il s'agit en effet de les administrer. Le logiciel fournit différentes solutions pour gérer les données et garantir leur qualité : l'attribution de rôles et de permissions, l'utilisation de listes de suivi et d'un système de veille ; l'implémentation de contraintes de qualité.

Premièrement, comme nous l'avons vu au cours du deuxième chapitre, le logiciel propose, à l'instar de ce qui se fait sur la base de connaissance Wikidata, d'attribuer différents rôles aux utilisateurs. Grâce à l'existence de groupes d'utilisateurs, comme par exemple le groupe des *administrateurs*, il est possible d'attribuer aux personnes faisant partie de ce groupe un ensemble de privilèges – appelés droits ou permissions –, tels que le fait de pouvoir créer une nouvelle propriété ou de bloquer un compte d'utilisateur. Cette fonctionnalité apparaît comme une opportunité pour le CegeSoma de clarifier sa politique de gestion et de contrôle des données d'autorité. En circonscrivant le rôle des différents acteurs concernés par l'édition et le contrôle des données d'autorité, l'institution se dote ainsi d'outils pour agir proactivement et limiter les problèmes de qualité, tels que ceux qui avaient impacté son thésaurus en l'absence de politique claire de gestion des métadonnées.

Dans le contexte d'une instance Wikibase, voici une vue synthétique des principales permissions pouvant être accordées :

— voir le contenu de la Wikibase
— créer un compte
— créer ou modifier des groupes et des droits d'utilisateurs
— bloquer un utilisateur
— créer, éditer ou protéger [37] une page [38]
— créer une page de discussion
— créer, modifier ou supprimer un élément [39] ou une propriété [40]
— fusionner deux éléments
— importer des lots de données (robot).

Concrètement, l'assignation de rôles – qui se fait en intégrant l'utilisateur à un ou plusieurs groupes d'utilisateurs – ne nécessite que quelques clics [41]

37. Cela signifie que des utilisateurs ne jouissant pas du même rôle et des mêmes permissions ne pourront pas la modifier.
38. Contenu *méta*, texte rédigé.
39. Données structurées.
40. Étrangement, aucun élément de la liste des permissions octroyées dans le cadre de Wikidata (voir https://www.wikidata.org/wiki/Wikidata:User_access_levels/fr) ne concerne spécifiquement la modification d'une propriété. Cela semble être une permission par défaut, étant donné que même sans être identifié sur Wikidata, il est possible d'ajouter ou de modifier des déclarations liées à une propriété. Voir par exemple : P1971|Nombre d'enfants https://www.wikidata.org/wiki/Property:P1971.
41. Par le biais de cette page : https://web.archive.org/web/20200713094240/https://adochs.arch.be/wiki/Special:UserRights.

et la modification des permissions attribuées à ces groupes [42] peut être configurée de façon aisée au sein du fichier de configuration *LocalSettings* [43], en revanche, cela requiert une réflexion en amont sur la façon dont l'institution souhaite procéder. En effet, il faut prendre en considération le fait qu'il y a une large marge de manœuvre, entre l'utilisation d'un nombre très restreint de groupes, avec une scission très nette entre des utilisateurs aux droits très limités et des administrateurs tout-puissants, et l'élaboration d'un système beaucoup plus pointu, jouant avec la granularité permise par les *espaces de noms* [44] que propose MediaWiki. Ainsi, le fondateur de l'instance Personaldata.io suggère par exemple qu'il serait intéressant d'attribuer des droits différents en fonction de ces espaces de noms. Il explique que Wikibase fournit ainsi un outil pour segmenter des domaines d'expertise, qui permettrait par exemple que seuls des experts des questions juridiques puissent modifier les pages associées à de tels contenus (PersonalData.io, 2020).

Dans le cadre du CegeSoma, la stratégie envisagée en concertation avec la responsable de l'accès numérique aux collections consiste à partir de l'existant pour ensuite déterminer quels rôles pourraient correspondre aux besoins relevés [45]. Concrètement, quatre types de profil ont été discernés [46] :

— les éditeurs (à savoir, les bénévoles, stagiaires, étudiants et chercheurs amenés à travailler sur les données d'autorité du CegeSoma)
— les experts (c'est-à-dire les membres de l'équipe scientifique du CegeSoma désignés pour effectuer un contrôle de qualité axé sur la cohérence et la rigueur des données du point de vue du fond)
— les administrateurs (soit les personnes responsables du bon fonctionnement de la Wikibase : de la gestion des éventuels spams à l'utilisation cohérente des propriétés et qualificatifs, en passant par la fusion d'éléments)
— les robots (il s'agit de comptes destinés à effectuer des modifications massives de façon semi-automatisée).

42. La liste actuelle des droits des groupes d'utilisateurs de notre Wikibase peut être consultée sur : https://web.archive.org/web/20200713094108/https://adochs.arch.be/wiki/Special:ListGroupRights, tandis que la liste des droits par défaut peut être consultée dans ce fichier de configuration, à partir de la ligne 5 077 : https://phabricator.wikimedia.org/source/mediawiki/browse/REL1_33/includes/DefaultSettings.php.

43. Voir : https://www.mediawiki.org/wiki/Manual:User_rights.

44. Il s'agit d'un sous-groupe de pages au sein d'un Wiki, concernant un sujet similaire. Par exemple, une Wikibase sera équipée par défaut d'un *espace de nom* pour les pages d'aide, d'un autre espace pour les pages d'utilisateurs, et ainsi de suite. Voir https://www.mediawiki.org/wiki/Manual:Namespace/fr.

45. Plutôt que de partir de la longue liste de permissions utilisées par exemple dans le cadre Wikidata : https://www.wikidata.org/wiki/Wikidata:User_access_levels/fr#Tableau.

46. Auquel s'ajoute un groupe d'utilisateurs par défaut, à savoir tout visiteur non enregistré de passage sur l'instance Wikibase.

Ce canevas de base, composé de catégories non exclusives, est destiné à être éprouvé et affiné avec le temps, quitte à inclure à l'avenir de nouveaux groupes d'utilisateurs si le besoin s'en fait sentir. Par ailleurs, la stratégie envisagée implique également de complémenter cette attribution des permissions par la création d'un groupe de travail spécialisé dans le contrôle qualité, constitué de membres de l'équipe scientifique du CegeSoma et amené à se réunir régulièrement [47]. Son objectif serait de pouvoir évaluer le fonctionnement en cours, aborder les problèmes rencontrés et enfin débattre de questions liées à la modélisation des données, pouvant par exemple conduire à la création de nouvelles propriétés. Si une page dédiée de la Wikibase pourrait être destinée à la soumission de nouvelles propositions – à l'instar de ce qui se fait sur Wikidata –, le maintien de réunions *de visu* semble en effet préférable dans le cadre des débats sur la pertinence de ces nouvelles propriétés. Cette configuration semble plus judicieuse [48] dans la mesure où le personnel du CegeSoma doit déjà s'adapter à beaucoup de nouveaux ajustements dans le cadre de l'utilisation de cette Wikibase et qu'il n'est pas coutumier de la tenue de débats en ligne comme le sont les membres de la communauté Wikidata.

Deuxièmement, le logiciel Wikibase permet de bénéficier de listes de suivi et d'un dispositif de veille appelé *patrolling*. Si les listes de suivi sont personnelles et destinées à pouvoir suivre les modifications apportées à une série d'éléments en particulier – par exemple des éléments créés par la personne elle-même –, le *patrolling* est quant à lui un outil collectif de contrôle de qualité des données. Ce terme désigne l'action de vérification et de validation de contenu par un *patrouilleur*, qui a pour mission d'annuler ou de rectifier les modifications non pertinentes effectuées par d'autres utilisateurs. Cela concerne donc tant le *spamming* et le vandalisme que des maladresses, incohérences ou inexactitudes. Lorsqu'un *patrouilleur* a examiné une modification, il peut la valider : elle sera alors indiquée comme ayant fait l'objet d'une *patrouille*, ce qui permettra aux autres personnes chargées du contrôle de qualité de savoir que cette modification a déjà été vérifiée et d'ainsi optimiser les opérations de contrôle de qualité [49].

47. Des premières démarches ont été effectuées en ce sens au mois de juin 2020 et il est prévu que des efforts concernant le contrôle qualité soient poursuivis au cours des prochains mois, quelle que soit l'issue de la Wikibase créée pour cette étude de cas.

48. À condition que des comptes rendus soient publiés sur la Wikibase à l'issue de ces réunions afin que toute la documentation demeure dans un seul et même espace.

49. « This allows people to coordinate their patrolling activity, such that all edits get checked at least once, but with less wasted effort (multiple people checking the same edit) » (Wikidata, 2020).

Le CegeSoma pourrait ainsi adopter un workflow – majoritairement [50] inspiré des pratiques de la communauté Wikidata (Wikidata, 2020) – composé de cinq étapes.

1. Passer en revue les dernières modifications de données [51] : identifier les modifications devant être supprimées [52] et les modifications devant être améliorées [53].
2. Obtenir des précisions (le cas échéant) : dans le cadre de données complexes ou ambigües, il peut être utile d'entrer en contact direct avec l'auteur des modifications.
3. Effectuer les opérations nécessaires : à savoir améliorer, supprimer ou marquer comme *ayant été patrouillé* tout ce qui doit l'être.
4. Prévenir les éditeurs (le cas échéant) : par exemple, si une personne est à l'origine d'erreurs récurrentes, elle doit recevoir les informations nécessaires pour éviter que le problème ne se reproduise.
5. Vérifier les autres contributions de l'utilisateur (le cas échéant) : l'auteur d'une certaine erreur ou imprécision est susceptible de l'avoir commise à plus d'une reprise.

Deux remarques doivent toutefois être énoncées. Premièrement, il est clair que dans le cadre d'importation automatisée de jeux de données contenant plusieurs centaines ou milliers de lignes, le contrôle qualité gagnera à être effectué en amont. Deuxièmement, il faut noter que dans le cadre de Wikidata, qui se caractérise notamment par le fait que ses utilisateurs sont issus des quatre coins du monde, tout ceci se fait exclusivement en ligne. Par exemple, la quatrième étape, qui implique une communication entre le patrouilleur et un autre utilisateur, se fait en postant un message directement sur la page personnelle – et publiquement accessible – de l'utilisateur. Dans le cas d'une situation comme celle du CegeSoma, où le personnel et les bénévoles sont amenés à se croiser entre les murs de l'institution, la question sera de voir dans quelle mesure les utilisateurs sont enclins à utiliser de façon systématique les outils associés à l'instance Wikibase. Si du point de vue de la maintenance des données, il est en effet crucial que tout puisse être

50. Nous avons ajouté l'étape numéro deux.

51. Il faut mentionner ici deux développements récents ou à venir. Premièrement, il existe depuis 2019 une application de *patrolling* destinée aux téléphones mobiles, *Speed Patrolling*, qui vise à simplifier les principales tâches de validation de contenu (voir la documentation : https://www.wikidata.org/wiki/User:Lucas_Werkmeister/SpeedPatrolling). Deuxièment, des efforts ont été initiés pour inclure un *ranking* permettant de prioriser les modifications devant être *patrouillées* en fonction du taux d'usage de l'élément correspondant (voir :*Rank wikidata changes for patrolling by usage* : https://phabricator.wikimedia.org/T173121).

52. Par exemple dans le cadre de vandalisme ou d'éléments ne répondant pas aux standards établis par le CegeSoma.

53. Par exemple dans le cadre de valeurs incorrectes.

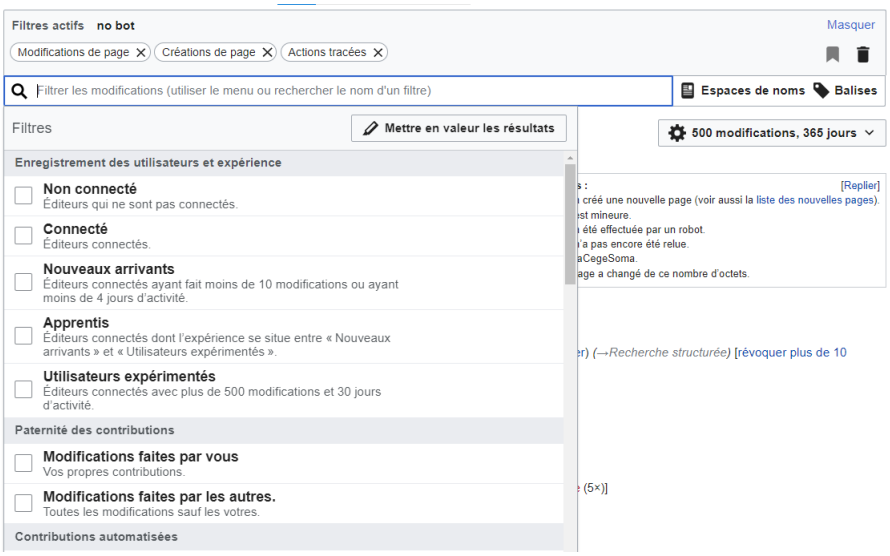


FIGURE 5.4 – Exemple de filtres permettant de cibler les modifications récentes à afficher dans le cadre d'une instance Wikibase. Source : Wikibase Adochs (https://adochs.arch.be/wiki/Special:RecentChanges).

consigné de façon permanente et transparente, il faut se rendre compte de la rupture que cela implique dans les méthodes de travail et de la résistance au changement que cela pourrait générer.

Outre ce système de *patrolling*, Wikidata propose une interface d'affichage des modifications récentes pouvant être personnalisée à l'aide de toute une série de filtres [54] utiles dans le cadre du contrôle de qualité des données. Ainsi, la figure 5.5 donne à voir un extrait des filtres pouvant être appliqués [55], comme le filtre *nouveaux arrivants* [56]. Ces filtres peuvent également servir pour mettre en surbrillance [57] les modifications récentes à l'aide de codes couleurs, en fonction de leurs caractérsitiques, mais plutôt que de décrire ces possibilités de façon exhaustive, l'essentiel ici est surtout de mettre en lumière le fait que l'infrastructure Wikibase est équipée de fonctionnalités beaucoup plus riches et fines que ne l'était par exemple le logiciel Pallas – uti-

54. Voir : https://www.mediawiki.org/wiki/Help:New_filters_for_edit_review.

55. Par ailleurs, une investigation plus poussée mériterait d'être menée afin de voir si certains filtres plus expérimentaux de Wikidata – basés sur des prédictions probabilistiques issues d'un service de *machine learning* et utilisés notamment dans le cadre de la lutte contre le vandalisme – pourraient également être implémentés dans le cadre d'une instance Wikibase ; voir : https://www.mediawiki.org/wiki/Help:New_filters_for_edit_review/Quality_and_Intent_Filters.

56. Qui permet d'afficher uniquement les modifications récentes ayant été effectuées par des utilisateurs à l'origine de moins de dix modifications ou ayant à leur compteur moins de quatre jours d'activité.

57. Voir : https://www.mediawiki.org/wiki/Help:New_filters_for_edit_review/Highlighting_function.

FIGURE 5.5 – Extrait de l'historique des révisions de Q10|Andrée de Jongh. Source : Wikibase Adochs (https://adochs.arch.be/w/index.php?title=Item:Q10&action=history).

lisé par le CegeSoma jusqu'il y a peu. En effet, ce dernier ne possède pas de module de contrôle en qualité en tant que tel. Le personnel soucieux de vérifer la qualité des données encodées se retrouvait ainsi contraint à faire appel à un administrateur système ou à effectuer des modifications *sur le tas*, lors de constats aléatoires d'erreurs ou en consultant les dernières descriptions, sans qu'une vue filtrée ne soit possible et sans que ne soit affiché le nom de la personne ayant encodé ces données [58]. Or, bénéficier de cette information peut être crucial, par exemple dans le cadre de l'amélioration de données imprécises ou contradictoires. La figure 5.5, montre le contraste que représente à ce niveau l'historique exhaustif et publiquement accessible associé à chaque élément Wikibase – ici l'extrait concerne l'élement Q10|Andrée de Jongh [59].

Enfin, il est possible de limiter les erreurs en restreignant l'usage de certaines propriétés à l'aide de contraintes de qualité issues de Wikidata. Il en existe une vingtaine, de nature différente. Par exemple, la contrainte d'utilisation unique par élément précise que « les éléments ne devraient jamais avoir plus d'une déclaration de cette propriété », c'est le cas par exemple pour les lieux de naissance ou de décès (Wikidata, 2020). Une autre contrainte populaire sur Wikidata est la contrainte de valeur distincte (sur Wikidata), qui précise que « chaque valeur devrait n'être présente [que] dans une seule déclaration de cette propriété dans tou[t] Wikidata » (Wi-

58. Entretien avec la responsable de l'accès numérique aux collections, 04.03.2020.
59. Voir : https://adochs.arch.be/w/index.php?title=Item:Q10&action=history.

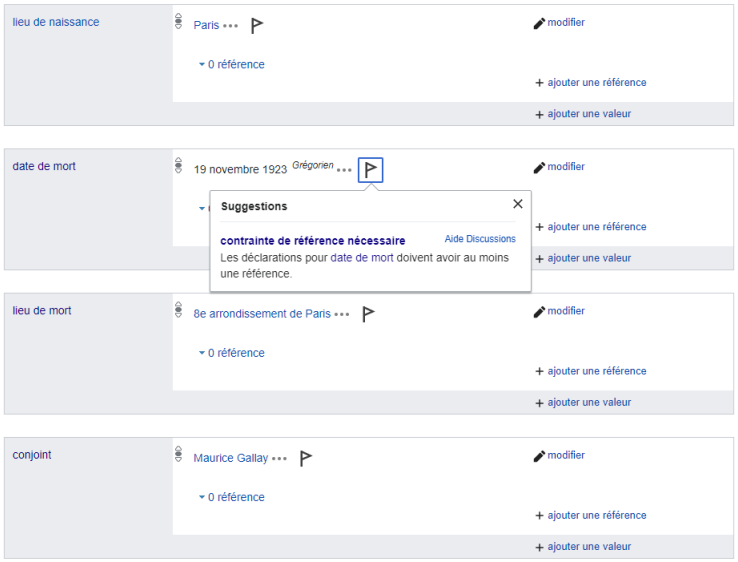


FIGURE 5.6 – Exemple de violations de contraintes de propriétés rencontrées sur la fiche Wikidata de l'élément Q67158249|Valentine Honegeer. Source : Wikidata (https://www.wikidata.org/wiki/Q67158249).

kidata, 2020), c'est le cas par exemple des identifiants externes tels l'identifiant ISNI ou l'identifiant SNAC. Un troisième exemple de contrainte fréquemment utilisée est la contrainte de type, qui permet de préciser de quelle nature devraient être « les éléments qui sont sujet d'une déclaration de cette propriété » ; par exemple les éléments décrits à l'aide d'une déclaration basée sur la propriété date de naissance devraient être de type Q5|être humain ou Q729|animal (Wikidata, 2020). Enfin, dans le contexte des données d'autorité, la contrainte de référence nécessaire semble particulièrement propice : elle stipule que les déclarations issues de cette propriété devraient être étayées par au moins une référence (Wikidata, 2020). La figure montre un exemple de la façon dont une violation de cette contrainte se manifeste sur Wikidata. Nous voyons ainsi que l'élément Q67158249|Valentine Honegger[60] a été décrit à l'aide de propriétés comme P19|lieu de naissance, P570|date de mort, P20|lieu de mort et P26|conjoint, qui requièrent toutes que la valeur encodée soit justifiée à l'aide d'une référence. Or ce n'est pas le cas actuellement, comme le signale le petit drapeau indiquant « quelques suggestions pour améliorer cette déclaration ».

À première vue, ce type de contraintes représenterait une plus-value significative pour l'instance Wikibase dédiée aux données d'autorité du CegeSoma. Elle pourrait en effet venir soutenir de manière très pragmatique la politique de contrôle qualité des données de l'institution. Malheureuse-

60. https://www.wikidata.org/wiki/Q67158249.

ment, cette extension [61] n'a pas encore pu être testée sur notre instance. En effet, son installation s'est avérée problématique, sans doute en raison d'un problème de compatibilité entre la version MediaWiki associée à notre installation Wikibase-docker et la version associée à l'extension WikibaseQualityConstraints [62]. Il s'agit donc d'une zone d'expérimentations qui mériterait d'être testée en priorité à l'avenir [63], étant donné son rôle stratégique dans le cadre de l'amélioration de la qualité des données. Et ce d'autant plus que le logiciel ne permet pas de restreindre lors de l'encodage le choix de valeurs possibles pour une certaine propriété.

Afin de tester les possibilités de cette instance nouvellement créée, la prochaine section propose d'approfondir quatre cas d'usage : recherche et accès aux données ; réutilisation des données ; exploration du potentiel des requêtes SPARQL fédérées pour compléter les informations contenues dans l'instance Wikibase en les enrichissant de données externes ; réconciliation de noms de personnes à partir des données contenues dans la Wikibase.

5.2 Cas d'usages

5.2.1 Recherche et accès aux données

Différents modes se présentent pour interroger la Wikibase et accéder aux données :

1. Recherche simple ou avancée
2. Recherche par le biais de l'API
3. Requêtes structurées (SPARQL)
4. Accès à l'ensemble des données (formats structurés)

Premièrement, la recherche simple permet, en utilisant l'interface graphique, de commencer à taper les premières lettres d'un mot pour que le système d'autosuggestion nous propose le(s) élément(s) correspondant(s) ; mais attention, cela fonctionne seulement à condition que le nom ait été encodé dans la langue sélectionnée dans l'interface. Ainsi, comme le montre la figure 5.7, l'élément Q3617|personne [64] n'a par exemple pas encore été

61. https://github.com/wikimedia/mediawiki-extensions-WikibaseQualityConstraints.
62. De la même façon que le décrit ce rapport de bug : https://phabricator.wikimedia.org/T197587.
63. Il serait également intéressant de s'intéresser à une solution alternative, à savoir les possibilités offertes par les Shape Expressions (ShEx), ce langage de validation de graphes RDF, voir par exemple le WikiProject Schemas : https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas ou l'outil WikiShape : http://wikishape.weso.es/. Une alternative serait d'utiliser la spécification du W3C, SHACL, par le biais d'un outil comme SHACL Play (voir https://shacl-play.sparna.fr/play/).
64. https://adochs.arch.be/wiki/Item:Q3617.

FIGURE 5.7 – L'élément Q3617|Personne n'a pas encore été traduit en anglais, néerlandais ou allemand. Source : Wikibase Adochs : https://adochs.arch.be/wiki/Item:Q3617.

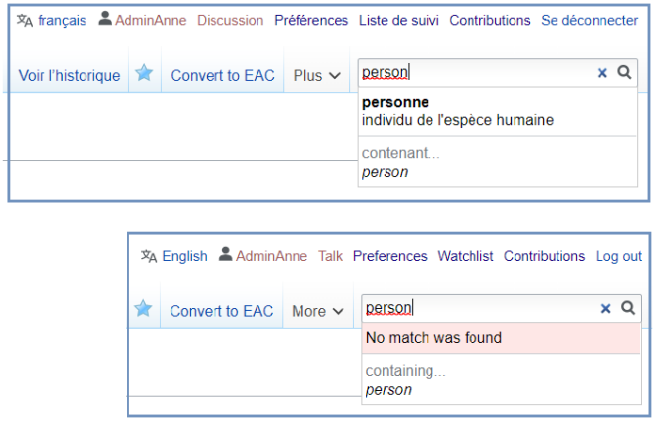


FIGURE 5.8 – Pas de résultat *(no match was found)* en tapant *person*... lorsque l'interface est en anglais. Source : Wikibase Adochs (https://adochs.arch.be/).

traduit en anglais. Cela signifie que si la recherche est effectuée sur une interface configurée en anglais, *person* ne renverra aucun résultat, comme l'illustre la figure 5.8. Cet exemple révèle l'impact du taux de traduction des données sur l'accès à ces dernières, ce qui renvoie à un autre type de contrôle et de maintenance à prendre en compte.

Par ailleurs, il faut noter que, par défaut, l'autocomplétion ne s'active pas si les premières lettres du libellé ne sont pas tapées de façon stricte. Par exemple, la recherche *de jongh* se solde par un échec (« aucune correspondance trouvée » (car le début du libellé ne correspond pas à *andrée de jongh*) [65]. Il s'agit toutefois d'une situation susceptible d'être améliorée à

65. Il faut insister et lancer une recherche pour des termes *contenant* ces lettres pour que la liste de résultats inclue l'ensemble des éléments pouvant coïncider.

l'avenir : comme le prouve un test effectué sur Wikidata avec une requête similaire, il est possible de configurer le moteur de recherche [66] de manière à ce que plusieurs autosuggestions soient proposées, quelle que soit la place dans le libellé des lettres encodées.

En revanche, les alias sont correctement pris en compte et permettent de retrouver l'entité correspondante en utilisant des formes alternatives plutôt que le libellé. Cela s'avère particulièrement utile, étant donné que le mode de recherche par défaut n'inclut pas d'algorithme de *fuzzy matching* : ainsi, l'élément portant le libellé « Paul **Henry** de la Lindy » ne sera pas retrouvé si c'est « Paul **Henri** de la Lindy » qui est encodé lors de la recherche et qu'aucun alias n'a encore été ajouté.

De plus, il faut noter l'importance des descriptions : ce sont elles qui permettent de réduire l'ambiguïté lorsque plusieurs éléments possèdent le même libellé. Or, si aucune description n'a été encodée, rien ne permet de distinguer deux *Jeanne Dupont*. Une solution *ad hoc* a cependant été déployée dans le cadre de Wikidata : il s'agit du gagdget *autodescription* développé par Magnus Manske [67]. Lorsqu'aucune description n'est disponible, comme c'est le cas par exemple pour Q3371461|Paul Henry de la Lindi [68] (voir figure 5.9), il est possible de générer une description automatique déployée à partir des déclarations qui ont été faites sur cet élément, comme le montre la figure 5.10, basée sur Wikidata. Ce système pourrait potentiellement être réutilisé dans le cadre d'une instance Wikibase tierce, cela nécessiterait toutefois l'adaptation d'un système de règles, afin d'identifier les propriétés-clés associées à une personne, telles que sa nationalité ou son occupation.

En complément de cette *search box*, il est également possible d'effectuer une recherche avancée [69], permettant par exemple de chercher une propriété plutôt qu'un élément, ou de rechercher parmi les pages de documentation en particulier. De plus, il faut noter l'existence de fonctionnalités pouvant être utiles pour la maintenance des données, comme cette recherche permettant d'obtenir la liste de tous les éléments ne possédant pas encore de libellé dans une certaine langue [70].

66. Concrètement, cela requiert d'installer l'extension WikibaseCirrusSearch (https://www.mediawiki.org/wiki/Extension:WikibaseCirrusSearch) que nous n'avons pas encore eu l'occasion de tester.
67. Voir : http://magnusmanske.de/wordpress/?p=64.
68. https://www.wikidata.org/wiki/Q3371461.
69. Voir : https://adochs.arch.be/w/index.php?title=Special:Search&profile=advanced&search=&fulltext=1.
70. Par exemple tous les éléments ne possédant pas de libellé en anglais, voir : https://adochs.arch.be/wiki/Special:EntitiesWithoutLabel?language=en&type=item.

FIGURE 5.9 – Il n'existe pas encore de description en français pour l'élément Q3371461|Paul Henry de la Lindi sur Wikidata. Source : Wikidata (https://www.wikidata.org/wiki/Q3371461).

FIGURE 5.10 – Exemple d'utilisation du gadget *Autodesc* dans le cadre de Wikidata. Source : Wikidata.

Une fois l'élément recherché identifié, il est naturellement possible de consulter la page [71] de l'élément – destinée à être lue ou modifiée par des humains –, mais également de récupérer le contenu de cette page dans un format lisible par des machines, grâce à un URI déréférençable (Wikidata, 2020a). En effet, l'*URI de concept* de Andrée de Jongh [72], qui correspond « à la vraie personne, et pas à sa description dans Wikidata [ou Wikibase] » (Wikidata, 2020a), permet, grâce à un processus de négociation de contenu, de déterminer le format dans lequel les données doivent être renvoyées. Outre une redirection vers la page HTML déjà mentionnée et destinée à une lecture humaine [73], il est ainsi possible de récupérer ces données dans un format structuré, lisible par des machines. Les clients du Web de données pourront ainsi recevoir les données au format JSON ou RDF en utilisant le code HTTP *Accept* (Wikidata, 2020a), mais cela peut également être réalisé via un navigateur web classique, en complétant l'URI à l'aide d'une extension précisant le format, comme : *.nt, .rdf, .ttl, ou .json* [74].

71. Par exemple, pour Q10|Andrée de Jongh : https://adochs.arch.be/wiki/Item:Q10.
72. À savoir : https://adochs.arch.be/entity/Q10.
73. À savoir : https://adochs.arch.be/wiki/Item:Q10.
74. Voir par exemple : https://adochs.arch.be/wiki/Special:EntityData/Q10.json.

Deuxièmement, un autre mode de recherche et d'accès aux données est proposé par le biais de l'API MediaWiki [75]. Il faut noter que l'API offre des possibilités de recherche plus limitées que le point d'accès SPARQL – qui permet de formuler des requêtes plus complexes, après la conversion des données en RDF – et son usage est, par défaut, soumis à des limites [76]. Cependant, les modules *wbgetentities* et *wbsearchentities* permettent toutefois d'accéder plus facilement au *JSON canonique* des pages d'entités (Wikidata, 2020a), ce qui peut être utile en cas de volonté d'obtenir toutes les informations disponibles sur une entité en particulier. Comme le montrent les exemples développés en annexes [77], il est possible de lancer un appel à l'API soit en effectuant une recherche basée sur une chaîne de caractères *(wbsearchentities)*, soit en utilisant directement l'identifiant numérique d'un élément en particulier *(wbgetentities)*. Précisons que manipuler les données via l'API requiert toutefois de se familiariser avec les aspects les plus techniques du modèle de données Wikibase et plus précisément les *snaks* (Baskauf, 2019), c'est-à-dire les combinaisons *propriétés - valeurs* [78] au cœur des déclarations et des qualificatifs faisant partie de ces déclarations (Wikidata, 2020c).

Troisièmement, il est possible d'accéder aux données à l'aide du langage de requête du Web sémantique, SPARQL. En effet, toute nouvelle instance Wikibase est équipée d'un point d'accès SPARQL, comme le propose Wikidata depuis 2015 (Wikidata, 2019). Ce mode d'accès permet de pleinement exploiter la richesse des données structurées. Ainsi, sous réserve de la disponibilité des informations [79], il est possible de formuler des requêtes plus complexes, pour par exemple se concentrer sur les membres de la Résistance dont on sait qu'ils étaient investis dans plusieurs réseaux ; observer les éventuels liens entre profession et type de reconnaissance nationale du statut de résistant ; retrouver toutes les femmes nées dans une même province et impliquées dans un réseau de résistance, analyser la distribution sur une carte des lieux d'arrestation de résistants ayant été fusillés ; ou encore extraire

75. https://adochs.arch.be/w/api.php.

76. Voir la documentation de l'API : https://www.mediawiki.org/wiki/API:Main_page.

77. Pour des détails concernant l'interrogation des données de la Wikibase à l'aide de requêtes SPARQL, consulter les annexes en ligne https://linkingthepast.org/query/ ou l'Annexe 9, page 384.

78. Sachant que la valeur peut également être une absence de valeur (*no value* ou *unkown value*).

79. Concrètement, les efforts de mise en place de cette Wikibase se sont concentrés sur un important travail de pré-traitement des données et sur la mise en place d'une structure à même d'optimiser la gestion des notices d'autorité pour des personnes, plutôt que sur la publication de données extrêmement détaillées au sujet de certains individus. Cependant, une fois cette infrastructure mise en place, il devient plus aisé de travailler sur l'enrichissement d'un sous-ensemble de notices, comme par exemple les membres d'un réseau de Résistance dont le CegeSoma possède les archives, tel que la Witte Brigade, et d'étoffer l'information les concernant à l'aide de nouvelles déclarations.

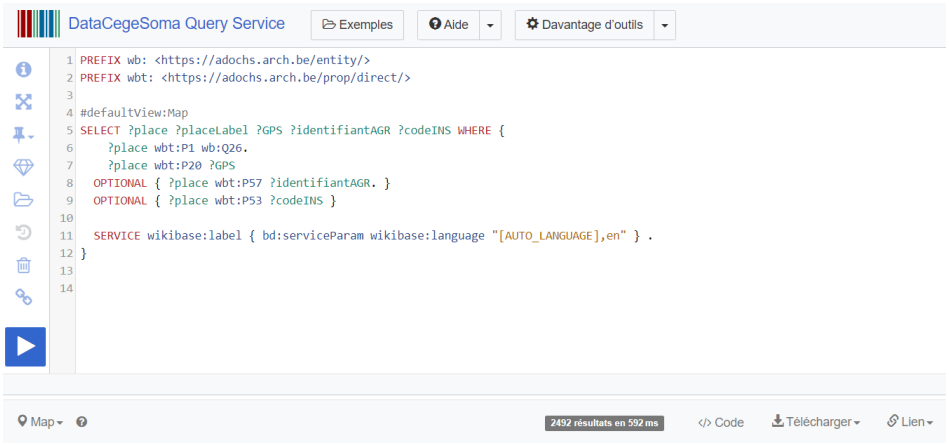


FIGURE 5.11 – Requête SPARQL visant à obtenir toutes les sections de communes de Belgique, accompagnées de leurs identifiants AGR, INS et Wikidata. Source : Wikibase Adochs (https://tinyurl.com/yyox4gle).

toutes les personnes ayant joué un rôle actif à la fois au cours de la Première et de la Seconde Guerre mondiale [80].

La consultation des données peut se faire de deux manières. D'une part, il est possible de passer par une interface web interactive [81] : ainsi, après l'encodage d'une requête basée par exemple sur la syntaxe SELECT, les résultats apparaissent au bas de l'écran sous forme de données tabulaires, ou potentiellement sous d'autres formes de visualisation. Les figures 5.11 et 5.12 montrent ainsi un exemple de requête et la représentation des résultats sur une carte.

Une fois ces lieux intégrés à la Wikibase sous forme d'éléments dotés de coordonnées GPS, il devient possible de croiser les données pour obtenir des visualisations plus pointues. Par exemple, la figure 5.13, montre les lieux d'exécution – par armes à feu – des personnes associées au fonds d'archives AA2346, à savoir des personnes ayant été impliquées dans des activités de résistance en Belgique au cours de la Seconde guerre mondiale et exécutées entre 1940 et 1944.

D'autre part, pour interroger le point d'accès SPARQL, il est également possible d'envoyer des requêtes HTTP utilisant les méthodes GET ou POST [82], cela signifie que des requêtes peuvent être adressées au point d'accès SPARQL de la Wikibase directement au sein d'un script Python. Ces données peuvent être ainsi facilement obtenues au format JSON, en vue d'être analysées ou

80. Ces exemples de questions sont issus d'échanges menés avec des historiens du CegeSoma au cours des mois de mai et juin 2019.
81. http://query-adochs.arch.be/.
82. Ces méthodes spéficient le type de demandes envoyées au serveur, pour plus de détails, voir https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/#query-operation.

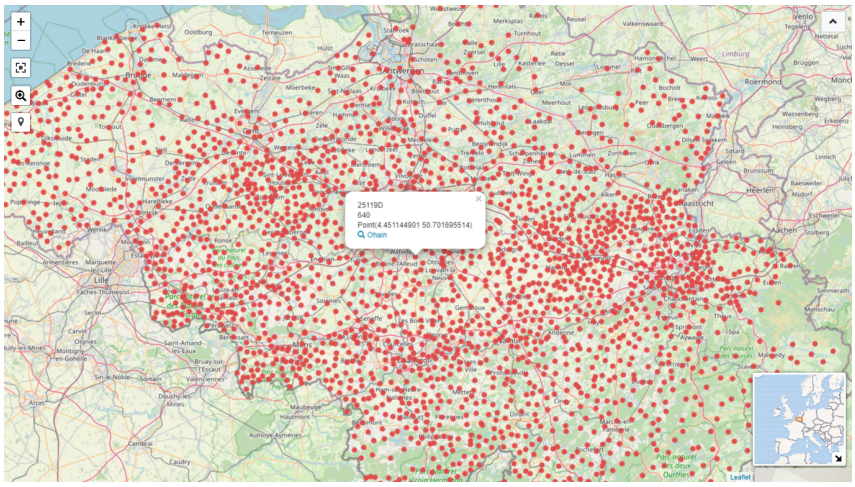


FIGURE 5.12 – Représentation sur une carte des résultats obtenus grâce à la requête montrée figure 5.11. Source : Wikibase Adochs (https://tinyurl.com/yyox4gle).

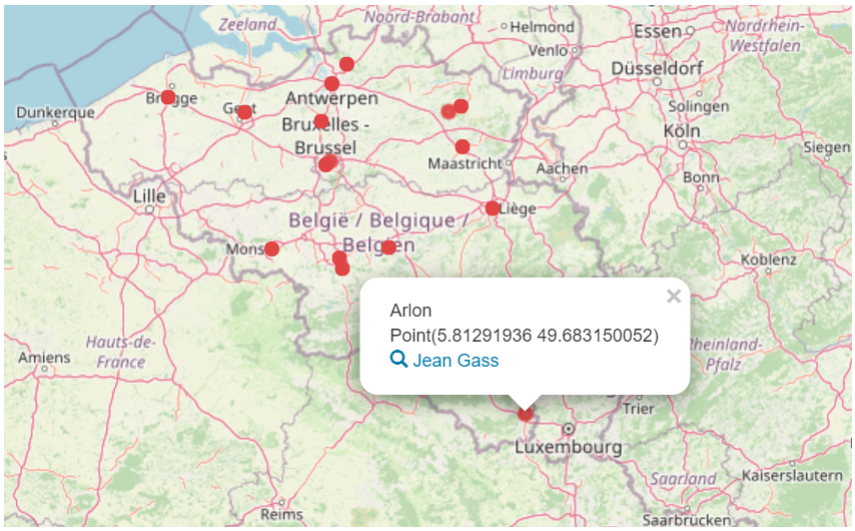


FIGURE 5.13 – Représentation sur une carte des lieux d'exécution des personnes associées au fonds d'archives AA2346, incluant Jean Gass, fusillé à Arlon le 07 juillet 1943. Source : Wikibase Adochs (https://tinyurl.com/y6bwyako).

réutilisées, par exemple dans le cadre d'un catalogue en ligne (voir section 5.2.2).

L'existence de ce point d'accès SPARQL signifie que l'institution peut fournir un accès privilégié à ses données structurées à toutes et tous, humains ou machines, qu'il s'agisse de chercheurs travaillant sur les collections du Centre, ou d'institutions tierces qui souhaiteraient récupérer des informations dans le cadre de requêtes fédérées (voir section 5.2.3). Des exemples de scripts contenant des requêtes SPARQL envoyées au point d'accès de la Wikibase sont présentés en annexes [83].

Quatrièmement, il est possible d'exporter des *dumps* de la base de connaissance permettant d'accéder à l'ensemble des données. Wikidata, qui crée et met à disposition une nouvelle version de façon hebdomadaire, recommande l'utilisation de dumps standards en JSON, chaque entité de la Wikibase (élément ou propriété) correspondant à un objet JSON individuel, placé sur une ligne distincte (Wikidata, 2020a). Bien que le processus tienne en une ligne de code [84], il faudrait évaluer dans quelle mesure cela rencontre un besoin dans le cadre de l'instance Wikibase du CegeSoma, au-delà de stratégies de sauvegarde des données, qui font appel à un autre type d'export [85].

5.2.2 Réutilisation

Comme l'explique Boydens, en concevant une application dans l'optique du concept WOPM *(Write Once Publish Many)*, il est possible de générer automatiquement des mêmes données structurées sous différents formats (Boydens, 2001). C'est le cas des instances Wikibase qui offrent une grande latitude dans la réutilisation des données. Ainsi, quel que soit le mode d'accès aux données (voir sous-section précédente), il est possible de les réutiliser au sein d'applications tierces, qu'il s'agisse d'un catalogue en ligne, d'un moteur de recherche, ou encore d'un service de génération de notices EAC CPF. Nous abordons dans cette sous-section ces trois cas de figure.

Le premier cas envisagé consiste à réutiliser des données Wikidata – ou d'une autre instance Wikibase – pour les afficher sur un catalogue ou une plateforme en ligne. Scott a par exemple réalisé une preuve de concept [86] afin d'intégrer des données issues de Wikidata et de Wikipedia sur le catalogue en ligne d'une bibliothèque universitaire, comme le montre la fi-

83. Consulter les annexes en ligne https://linkingthepast.org/query/ ou l'Annexe 9, page 384.

84. Voir : http://learningwikibase.com/install-wikibase/#exporting-data-as-a-json-or-rdf-dump.

85. Il s'agit d'effectuer un *dump* de la base de données en SQL, voir : http://learningwikibase.com/install-wikibase/#backing-up-data-from-docker-volumes.

86. Le code source est disponible en ligne : https://gitlab.com/denials/wikidata-music-infocard.

gure 5.14. Les données sont récupérées en temps réel à l'aide d'une requête SPARQL renvoyant les données au format JSON-LD, qui sont ensuite manipulées à l'aide de Javascript permettant leur affichage sur la page web concernée (Allison-Cassin et Scott, 2018).

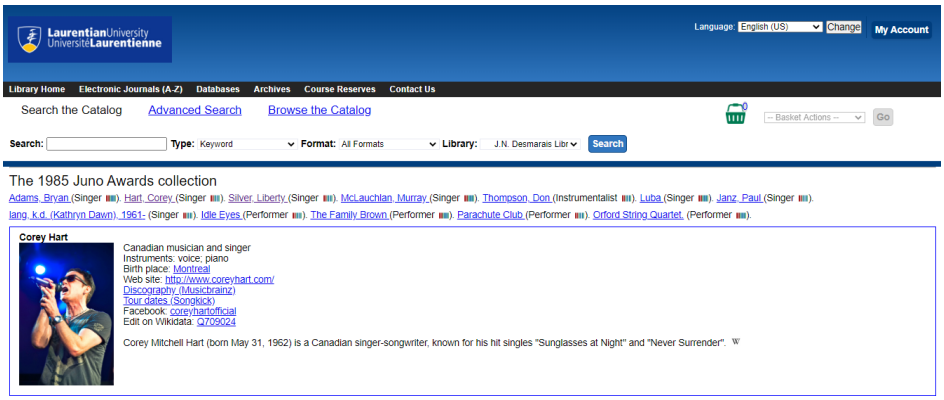


FIGURE 5.14 – Exemple d'*infocard* relative au chanteur Corey Hart, affichée sur le catalogue en ligne de la Laurentian University Library et créée en temps réel à partir des données de Wikidata et de Wikipedia. Source : Laurentian University Library (https://laurentian.concat.ca/eg/opac/record/349295).

Il est intéressant de noter que les données n'ont pas encore été réconciliées avec Wikidata : tout se fait au moment de l'envoi de la requête, comme l'expliquent Allison-Cassin et Scott :

> The SPARQL query [...] is implemented as a UNION of three different subqueries, to find musicians, bands, or musical ensembles with names or aliases that match the requested contributor's name, and to potentially return supplemental data to include in the infocard, along with a possible link to a corresponding Wikipedia entry[87]. (Allison-Cassin et Scott, 2018)

Dans le cadre du CegeSoma, un processus similaire pourrait par exemple être déployé pour enrichir les pages Personnalités de la plateforme de valorisation Belgium WWII[88]. Ces pages rédigées par des experts décrivent l'implication d'une centaine de personnes dans la Seconde guerre mondiale en Belgique, mais ne sont pas forcément dotées d'informations biographiques synthétiques permettant de situer la personne dans le temps et l'espace. La figure 5.15 donne un aperçu, à titre purement indicatif[89], de la façon dont

87. L'API MediaWiki est ensuite utilisée pour récupérer le texte d'introduction de l'article Wikipedia concerné.

88. https://www.belgiumwwii.be/.

89. Nous avons utilisé des outils de développement web permettant de modifier le code source de la page, et dès lors, le texte affiché, afin de simuler le résultat envisagé.

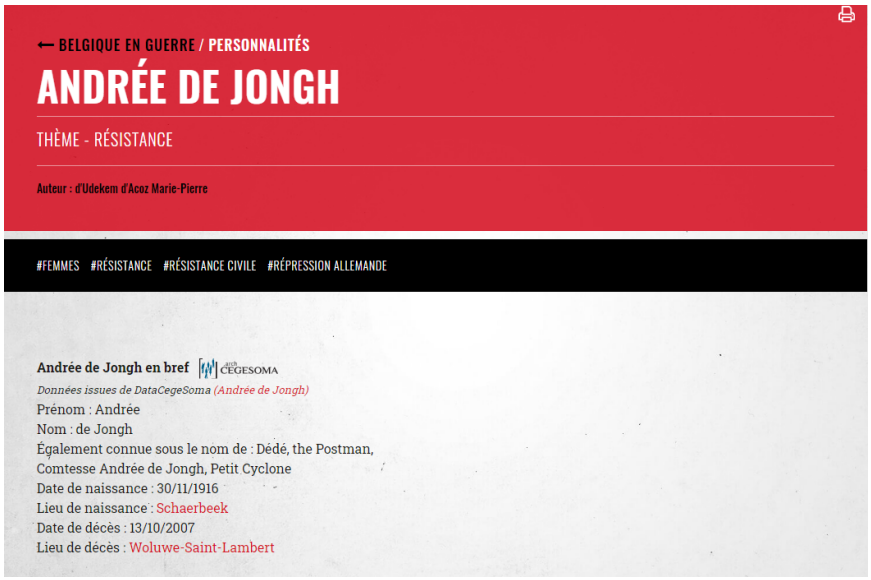


FIGURE 5.15 – Exemple (fictif) de fiche signalétique – ici pour Andrée de Jongh – qui serait intégrée sur la plateforme Belgium WWII et alimentée à l'aide de données extraites en temps réel de la Wikibase. Source : Belgium WWII (https://www.belgiumwwii.be/belgique-en-guerre/personnalites/andree-de-jongh.html).

des données structurées issues de la Wikibase pourraient être affichées en tête de page, reprenant des informations biographiques ciblées telles que les dates et les lieux de naissance et de décès.

Si l'évocation d'un tel projet a rencontré un écho favorable au sein de l'équipe scientifique du CegeSoma, il faudrait toutefois approfondir l'idée en discutant des données susceptibles d'être utiles aux lecteurs. Cela permettrait ainsi d'adapter le code issu du projet de la Laurentian University Library (Allison-Cassin et Scott, 2018) et de générer les requêtes SPARQL correspondantes. Il faut noter que si le code Javascript peut être configuré de manière à n'afficher que le contenu disponible [90], un travail supplémentaire serait néanmoins requis en amont, afin de décliner ces requêtes dans les trois langues de publication de Belgium WWII [91]. Cependant, l'effort initial consenti permet ensuite un important retour sur investissement : l'information ne doit désormais être modifiée plus qu'une seule fois, directement au sein de la Wikibase, à l'instar des modifications des infoboxes Wikipedia

90. Sachant que le degré de complétude des données de la Wikibase relatives aux personnalités décrites sur Belgium WWII est susceptible de varier, il est intéressant de bénéficier d'un système flexible permettant d'afficher le plus grand nombre d'informations possibles en fonction de leur disponibilité, plutôt que de devoir se restreindre au plus petit dénominateur commun.

91. Français, néerlandais et allemand.

qui sont désormais effectuées directement sur Wikidata et automatiquement synchronisées.

Le deuxième cas de figure vise à tirer profit des données structurées en les utilisant en coulisses d'un moteur de recherche. En effet, la sémantisation et la structuration de l'information – permettant désormais à une machine de comprendre que l'entité Andrée de Jongh correspond à une personne, de genre féminin, née en 1916, à Schaerbeek, impliquée dans la Seconde guerre mondiale – peut être mobilisée dans le cadre de requêtes d'utilisateurs. En effet, si le nom *Andrée de Jongh* a été tagué au sein d'un document [92] et associé à l'URI d'Andrée de Jongh, la machine *saura* que tel document concerne une femme née en 1916 en Belgique et impliquée dans la Seconde guerre mondiale. Cette connaissance peut par exemple être exploitée dans le cadre de recherches menées au sein de corpus de presse ancienne numérisée. Ainsi, au lieu d'effectuer une recherche simple basée sur une chaîne de caractères en particulier, l'utilisateur peut initier des recherches plus larges, non nominatives. Il pourrait par exemple rechercher tous les extraits d'un certain journal mentionnant des femmes nées à Schaerbeek, âgées de 20 ans ou plus lors du déclenchement de la Seconde Guerre mondiale, dont on sait [93] qu'elles ont été liées à ce conflit. Parmi les résultats figureraient vraisemblablement des extraits faisant mention de Andrée de Jongh, mais pas seulement, permettant ainsi au chercheur de *ratisser* plus large et de faire ainsi de nouvelles découvertes, dans le cadre d'une sérendipité *cadrée*. S'il s'agit encore d'une chimère à l'heure actuelle pour le CegeSoma, il s'avère qu'un tel dispositif a déjà été implémenté – de façon expérimentale – par la Bibliothèque nationale des Pays-Bas [94].

En effet, van Veen (2019) a illustré la puissance que permet l'utilisation conjointe d'entités nommées et de Wikidata dans le cadre de la presse ancienne numérisée à l'aide d'exemples. Il présente ainsi le scénario d'une recherche formulée en langage naturel automatiquement *traduite* en requête SPARQL afin de présenter les extraits de presse correspondants :

> A query string entered between square brackets, for example « [roman emperor] », is expanded by a *best guess* SPARQL query in Wikidata, in this case resulting in entities having the property « position held=roman emperor. ». These in turn are used to do a search for articles containing one or more mentions of a Roman emperor, even if the text *roman emperor* is not present in the article (van Veen, 2019, p. 74).

92. Ou au sein de la description de ce document.
93. C'est-à-dire que l'information est présente dans la base de connaissance utilisée comme ressource.
94. Voir : http://www.kbresearch.nl/xportal/#demo.

Mais l'interface développée par la Bibliothèque nationale des Pays-Bas permet également de formuler soi-même une requête SPARQL Wikidata. Nous avons ainsi testé une requête [95] visant à retrouver des extraits de presse concernant des femmes nées en Belgique, connues comme ayant été des membres de la Résistance belge. Les résultats [96] contiennent – sur la deuxième page de résultats – à juste titre la mention de l'entité Q6762930|Marie Louise Habets [97], infirmière et ancienne religieuse belge qui a été impliquée dans la Seconde Guerre mondiale comme résistante. En revanche, la qualité des résultats reste conditionnée par les difficultés propres à la reconnaissance d'entités nommées. Ainsi, parmi les résultats figure un extrait d'un journal de 1939 [98] enrichi à l'aide du tag *Marie-Louise Habets*, qui contient en réalité une liste de noms de personnes ayant réussi le MULO, un examen scolaire des Pays-Bas [99]. Si *M. Habets* figure parmi ces noms, il semble cependant fort peu probable que ce soit Marie Louise Habets qui ait passé cet examen en 1939 à Herrlen, dans la Province du Limbourg. En effet, d'après la page Wikipedia qui lui est dédiée [100], elle est devenue religieuse en 1926, partie en mission au Congo belge en 1933, avant de rentrer en Belgique en 1939, atteinte de la tuberculose. Si la Bibliothèque nationale des Pays-Bas est consciente des limites de cet enrichissement réalisé de façon automatisée à l'aide d'algorithmes et informe en toute transparence les utilisateurs de ce service web [101], il s'agit toutefois d'éléments que le CegeSoma doit garder à l'esprit, s'il décidait à l'avenir de bâtir un tel dispositif autour de sa plateforme en ligne *War Press*.

Le troisième cas de figure est né de besoins formulés par des membres de l'équipe scientifique du CegeSoma. L'objectif est de pouvoir générer un fichier EAC-CPF à partir des données de la Wikibase, afin de disposer d'un format de données adapté aux pratiques des Archives de l'État tout en limitant l'encodage d'informations redondantes sur différents supports. Il se trouve que cette idée a également fait l'objet d'un projet mené par l'équipe

95. À savoir : SELECT?p?pLabelWHERE{?pwdt:P21wd:Q6581072.?pwdt:P106wd:Q1397808.?pwdt:P19?place.?placewdt:P17wd:Q31.SERVICEwikibase:label{bd:serviceParamwikibase:language[AUTO_LANGUAGE],nl,en,de,fr.}}.

96. Accessibles en suivant ce raccourci : https://tinyurl.com/y665f59a.

97. https://www.wikidata.org/wiki/Q6762930.

98. Il s'agit du Limburger Koerier, et plus précisément de l'édition du 22 juillet 1939, voir https://resolver.kb.nl/resolve?urn=ddd:010985996:mpeg21:a0311.

99. Voir https://nl.wikipedia.org/wiki/Meer_uitgebreid_lager_onderwijs.

100. https://en.wikipedia.org/wiki/Marie_Louise_Habets

101. Nous pouvons ainsi lire : « This site represents experimental work in progress. The links for named entities to DBpedia and Wikidata are therefore not yet reliable (but you may correct them) ».

FIGURE 5.16 – Un bouton *convert to EAC* a été ajouté en modifiant le code Javascript de la page MediaWiki :Common.js. Source : Wikibase Adochs (https://adochs.arch.be/wiki/Item:Q10).

dont nous faisions partie [102] lors du premier hackathon organisé par les Archives nationales de France, en décembre 2018. Dans le cadre de la présente étude de cas, nous avons adapté [103] ce code – repris dans l'Annexe 10 [104] – à l'instance Wikibase accueillant les données du CegeSoma.

Les figures 5.16 et 5.17 illustrent comment un clic sur le bouton associé à une fiche Wikibase permet de générer un fichier EAC-CPF [105] reprenant les informations de base sur cette personne [106].

Enfin, ajoutons qu'il pourrait également être opportun de réutiliser des outils de visualisation de données déployés dans le cadre de Wikidata, c'est-à-dire des interfaces donnant à voir les données de Wikidata dans un format plus convivial et destiné au grand public : comme Reasonator [107], Crotos [108], Sciences stories [109] ou encore Histropedia [110]. S'il y a tout lieu de penser qu'ils pourraient être déclinés pour être utilisés dans le cadre d'une instance Wikibase, cela nécessiterait toutefois de plus amples investigations.

102. Il s'agit de l'équipe Wikilinki, qui a remporté le prix du *Coup de cœur sémantique*, voir : http://www.archives-nationales.culture.gouv.fr/resultats-du-hackathon-des-archives-nationales.
103. Avec la contribution d'un membre de l'équipe Wikilinki, Adrien Di Mascio.
104. Page 385.
105. Exemple pour Andrée de Jongh : http://eac-cpf.herokuapp.com/eac/Q10.
106. Une version plus aboutie nécessiterait d'affiner le travail d'alignement entre les propriétés Wikibase et les éléments de l'EAC-CPF, limité dans le cadre de cette démonstration à des éléments facilement assimilables, telles que les années de naissance et de décès.
107. https://reasonator.toolforge.org/.
108. https://zone47.com/crotos/.
109. http://www.sciencestories.io/.
110. http://www.histropedia.com/.

FIGURE 5.17 – Extrait du fichier xml EAC-CPF généré à partir de la fiche Wikibase dédiée à Andrée de Jongh. Source : http://eac-cpf.herokuapp.com/eac/Q10.

5.2.3 Requêtes fédérées

Le troisième cas d'usage concerne l'interrogation simultanée de plusieurs points d'accès SPARQL. En effet, la vision d'une gestion semi-centralisée des données d'autorité prend petit à petit forme grâce à l'existence des requêtes SPARQL fédérées : il est possible, du moment que l'on dispose de liens d'équivalence et des points d'accès SPARQL requis, de bénéficier de l'apports d'autres institutions. Concrètement, il s'agit d'utiliser le service de requête de la Wikibase comme pour une requête classique, puis d'utiliser une jointure, c'est-à-dire un pivot permettant de faire le lien entre les données de la Wikibase et les données d'une autre base de connaissance – Wikidata dans l'exemple qui suit. L'existence de telles requêtes pourrait ainsi permettre à l'institution de se concentrer sur les données d'autorité propres à son domaine de compétence (les questions de société et les grands conflits du XX^{e} siècle, dans le cas du CegeSoma), en tirant parti de données externes pour ce qui s'étend au-delà de ce périmètre.

À titre illustratif, nous avons par exemple testé [111] comment pouvaient être récupérées, pour une personne donnée, les éventuelles informations présentes sur Wikidata au sujet de ses occupation(s), affiliation(s) à un parti politique, distinction(s) reçue(s) ou au sujet des institutions possédant des archives à son sujet. La figure 5.18 montre le résultat de cette requête [112] utilisant l'identifiant Wikidata stocké sur la Wikibase comme pivot pour combi-

111. Pour le détail sur cette opération et d'autres exemples, consulter les annexes en ligne https://linkingthepast.org/feder/ ou l'Annexe 11, page 395.
112. https://tinyurl.com/y868bhgu.

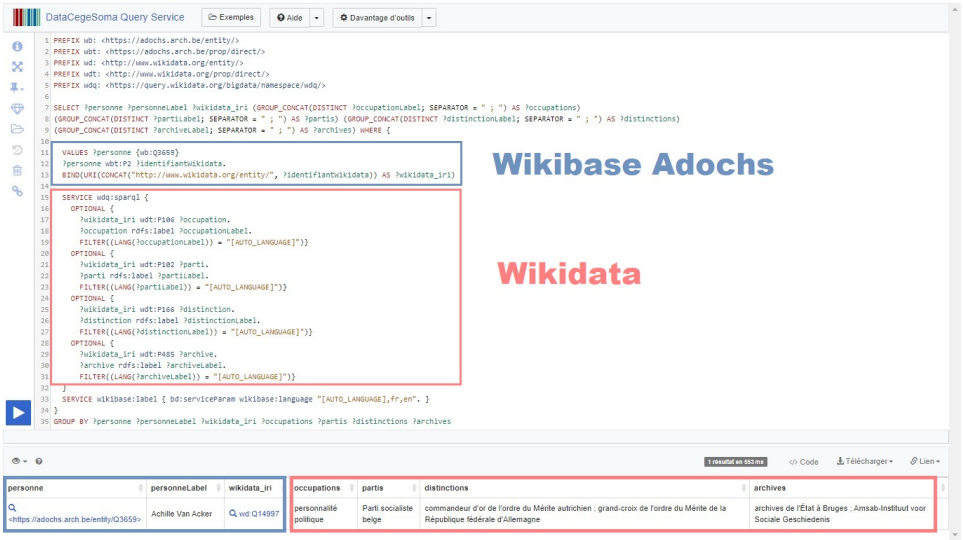


FIGURE 5.18 – Exemple de requête fédérée combinant des données issues d'une instance Wikibase et de Wikidata. Source : Wikibase Adochs (https://tinyurl.com/y868bhgu).

ner l'information issue de la Wikibase (en bleu) avec l'information issue de Wikidata (en rose).

L'existence de cette possibilité signifie qu'au lieu d'enrichir laborieusement les données du CegeSoma à l'aide de données externes avant leur importation dans la Wikibase (voir sous-section 4.2.4), il est possible de tirer avantage des requêtes fédérées pour récupérer l'information à la demande, en bénéficiant dès lors des dernières mises à jour des données.

Il faut toutefois prendre en considération le fait que cela requiert que des liens d'équivalence aient d'ores et déjà été établis entre les entités de la Wikibase et des entités externes, pour peu que ces dernières existent seulement ! D'autre part, il faut noter que la requête doit être adaptée à chaque fois aux informations souhaitées, en tenant compte de l'ontologie *cible* – là où des données enrichies a priori s'affichent par défaut sur les pages Wikibase des entités –, ce qui rajoute fatalement une couche de complexité pour l'utilisateur qui n'est peut-être pas familier du langage de requête SPARQL. Par ailleurs, ce genre de requêtes peut rapidement entraîner des problèmes de performance du serveur ou du service de requête *cible*, notamment s'il y a beaucoup de requêtes uniques réalisées en parallèle [113].

En dépit de ces limites, il est néanmoins possible pour le CegeSoma d'envisager l'utilisation de telles requêtes dans le cadre de plateformes ou cata-

113. Voir par exemple les limites du Wikidata Query Service : https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service.

logues en ligne, à l'instar de ce que propose la plateforme inventaire.io [114], qui est basée sur l'utilisation conjointe de données *locales* et de données issues de Wikidata (inventaire.io, 2020) [115]. De plus, comme nous l'avons déjà précisé, ce genre d'opérations sera peut-être amené à être facilité à l'avenir, dans le cadre des efforts déployés par l'équipe de développement derrière Wikibase pour développer les possibilités de fédération au sein de cet écosystème d'instances Wikibase. Enfin, il faut garder à l'esprit que pour que de tels mécanismes puissent se développer de façon optimale, il est dans l'intérêt de tous qu'un cercle vertueux puisse se mettre en place et que les institutions comme les chercheurs endossent la responsabilité de ne pas seulement tirer profit du potentiel de Wikidata, mais également de contribuer, selon leurs possibilités. Par exemple, cette requête [116] révèle que, pour l'instant, seules 50 femmes décrites comme étant de nationalité belge sont caractérisées sur Wikidata comme ayant été résistantes : cela laisse entrevoir le potentiel que revêtent les données d'autorité du CegeSoma pour Wikidata, notamment dans le cadre d'efforts pour réduire le *gender gap*. Cependant, d'autres formes de contribution, plus discrètes, telles que la correction d'erreurs et d'incohérences, sont également valorisées par la communauté [117].

5.2.4 Service de réconciliation

Un dernier cas d'usage auquel nous pouvons penser est la création d'un service de réconciliation basé sur l'API associée à notre Wikibase. Que cela soit dans le contexte de la détection de doublons avant l'importation de nouveaux jeux de données ou encore dans le cadre de la désambiguïsation d'entités nommées extraites de la presse ancienne numérisée [118], bénéficier d'un tel outil représenterait une réelle plus-value pour l'institution. Plus globalement, cela pourrait également permettre à d'autres utilisateurs ou d'autres institutions d'aligner plus facilement leurs données avec les autorités du CegeSoma.

114. Un service en ligne de prêt de livres entre individus, voir : https://inventaire.io/welcome.

115. Ces dernières sont utilisées en cache afin d'éviter les problèmes de performance : « le serveur fait la requête à Wikidata et en sauvegarde les résultats localement dans une base de données LevelDB [...] la liste des oeuvres ne sera donc pas mise à jour pendant 1 mois, sauf si demandé explicitement par un utilisateur. » [Messages Telegram] max lath 08.07.2020.

116. https://w.wiki/pBq.

117. Comme l'illustre par exemple l'un des objectifs de Wikimedia Deutschland pour 2020 *(Re-users of the data support Wikidata)*, qui vise à ce que les rapports d'erreurs et d'incohérences issus de la réutilisation de données Wikidata soient librement mis à disposition (Wikimedia Deutschland, 2019b).

118. Par exemple dans le cadre du projet The Belgian War Press (https://warpress.cegesoma.be/fr.

Par chance, le dépôt GitHub *openrefine-wikibase* annonce que le service de réconciliation OpenRefine développé initialement pour Wikidata a été adapté afin de pouvoir être désormais réutilisé dans le cadre de n'importe quelle instance Wikibase :

> This service can be configured to run against another Wikibase instance than Wikidata. The Wikibase instance will need to have an associated SPARQL Query Service, and some properties and items will need to be set up. (Delpeuch, 2020)

Dans le cadre de cette étude de cas, nous avons brièvement testé en local [119] le déploiement d'un tel service. Si les détails liés à la configuration de cet outil sont présentés en annexes [120], deux formes de difficultés doivent être soulignées ici. Premièrement, les noms comportant des accents ont suscité des problèmes au moment de la réconciliation, ce qui nécessiterait une investigation plus poussée [121]. Deuxièmement, les fonctionnalités de *fuzzy matching* implémentées dans le cadre du service de réconciliation Wikidata et permettant de trouver des équivalences approximatives, n'étaient pas opérationnelles lors de nos tests. Or, c'est précisément dans ce type de fonctionnalités que se loge la puissance d'un tel outil. Cependant, le fait que le développeur de ce service ait lui-même signalé ce problème [122] donne bon espoir sur le fait que ce problème pourrait être traité au cours des mois à venir.

Enfin, précisons que l'importance de ce cas d'usage basé sur l'utilisation d'un service de réconciliation est susceptible de se renforcer encore plus à l'avenir, étant donné que depuis le mois d'août 2020 [123], le logiciel OpenRefine peut maintenant également être utilisé pour importer des données directement sur n'importe quelle instance Wikibase. Cela signifie qu'il est désormais possible d'imaginer un workflow plus harmonieux évitant par exemple les aller-retours entre OpenRefine et Quick Statements.

Ce cinquième chapitre visait à présenter de manière très concrète comment les données d'autorité préparées au cours des étapes précédentes sont créées puis administrées au sein d'une installation Wikibase. Il avait également pour objectif d'esquisser différentes manières dont les données peuvent ensuite être utilisées. Cela nous a permis de montrer divers cas d'usages,

119. C'est-à-dire que le service tourne sur une machine individuelle plutôt que sur un serveur qui permettrait d'en faire un service web accessible à d'autres utilisateurs.
120. Consulter les annexes en ligne https://linkingthepast.org/recon/ ou l'Annexe 12, page 396.
121. Le problème a été signalé ici : https://github.com/wetneb/openrefine-wikibase/issues/82.
122. Voir : https://github.com/wetneb/openrefine-wikibase/issues/80.
123. Comme le révèle la clôture de cette *issue* GitHub intitulée *Extend Wikidata extension to support arbitrary Wikibase instances* : https://github.com/OpenRefine/OpenRefine/issues/1640.

qu'il s'agisse d'une simple recherche par nom, de modes de recherche plus avancés, de réutilisation des données dans le cadre d'une autre interface, de requêtes fédérées permettant d'interroger simultanément plusieurs sources, ou encore de la mise en place d'un système de réconciliation basé sur les données contenues dans la Wikibase. Cependant, si l'implémentation d'une instance Wikibase pour la gestion des données d'autorité offre indéniablement de nouvelles perspectives intéressantes au secteur culturel, certains bémols ne peuvent être passés sous silence. C'est ce que nous nous proposons de mettre en lumière au cours du chapitre suivant sous la forme d'une analyse SWOT (*Strengths, Weaknesses, Opportunities, Threats*) suivie de recommandations généralisables à d'autres cas que le CegeSoma.

6 | Analyse SWOT et recommandations

Introduction

Ce chapitre propose de discuter les observations et enseignements tirés de notre étude de cas, afin d'en dégager des recommandations généralisables. La construction de ce chapitre est inspirée des quatre paramètres examinés dans le cadre d'une analyse SWOT : *Strengths, Weaknesses, Opportunities, Threats*. Cette méthode d'analyse, dont les prémices remontent aux années 50 (Leigh, 2009), fournit un cadre de planification stratégique pour évaluer une organisation, un plan, un projet, une personne ou une activité commerciale (Gürel et Tat, 2017). Elle vise à considérer « the inhibitors and enhancers to performance that an organization encounters in both its internal and external environments » (Leigh, 2009, p. 1089). L'analyse SWOT permet ainsi de déterminer les actions à prendre (Gürel et Tat, 2017), en identifiant les forces internes et opportunités externes qu'une organisation peut exploiter pour atteindre ses objectifs, tout en cherchant à atténuer les faiblesses internes et menaces externes (Lewis and Littler, 1997, cité par Leigh, 2009).

Ce type d'analyse, également utilisé dans le contexte de la bibliothéconomie et des sciences de l'information, par exemple dans le cadre de l'émergence des médias sociaux (Fernandez, 2009) ou des MOOC (Kaushik, 2018), semble approprié à notre étude de cas dans la mesure où les paramètres de la matrice SWOT permettent d'analyser les possibilités et les limites d'une Wikibase pour CegeSoma, en distinguant les facteurs internes des facteurs externes. Ces facteurs, repris dans le tableau récapitulatif 6.1, sont détaillés au cours des quatre sections qui suivent, tandis que des recommandations sont émises dans une cinquième section venant clôturer ce chapitre et cette étude de cas.

Forces	Faiblesses
Savoir-faire issu du projet Adochs	Ressources limitées
Expérience dans des projets DH	Résistance au changement
Expertise historique	Données en silos
Collections déjà indexées	Données non standardisées
Stratégie d'ouverture des données	Données sommaires
Partenaires institutionnels	Données peu structurées
Partenaires académiques	Granularité de l'indexation

Opportunités	Menaces
Besoins des utilisateurs	Instabilité du logiciel
Nombreux atouts du logiciel	Courbe d'apprentissage
Politique Open Data	Sources externes de qual. variable
Positionnement stratégique	Maintenance de l'infrastructure
Projet de norme *Records in Contexts*	Multiples données à synchroniser
Jeux de données externes	Intégration aux workflows actuels
Synergies potentielles	Performance et scalabilité
Expertise technique des AGR	Pérennité des URI
Inclusion d'autres types de données	

TABLE 6.1 – Tableau récapitulatif de l'analyse SWOT destinée à évaluer la pertinence d'une Wikibase pour le CegeSoma.

6.1 Forces

Savoir-faire issu du projet Adochs
Le savoir-faire développé au cours de l'installation, de la configuration et de la personnalisation d'une instance Wikibase – dans le cadre du projet Adochs – représente un acquis sur lequel peut s'appuyer l'institution.

Expérience préalable du CegeSoma dans des projets DH
Au cours des dix dernières années, le Centre a activement contribué à

plusieurs projets s'inscrivant dans la lignée des *Digital Humanities* [1] et a ainsi acquis une précieuse expérience dans la gestion de tels projets.

Expertise historique

Le CegeSoma dispose d'une équipe scientifique permanente incluant des historiens spécialisés, entre autres, dans le parcours individuel des résistants ou des collaborateurs au cours de la Seconde Guerre mondiale [2] et faisant autorité en la matière.

Collections déjà indexées

L'institution a commencé il y a plus de 20 ans à indexer ses collections à l'aide d'un thésaurus incluant des noms propres : des milliers de noms de personnes sont donc d'ores et déjà associés aux fonds d'archives ou de photographies qu'elle conserve et n'attendent que d'être davantage valorisés.

Stratégie d'ouverture des données

L'institution accorde une importance croissante à l'accès numérique aux collections (CegeSoma, 2018a; Gillet, 2019) et notamment à la libre mise à disposition des données, une stratégie qui s'inscrit dans la politique *open data* encouragée par les Archives de l'État [3].

Partenaires institutionnels

Le Centre bénéficie d'un large réseau de relations qui ne cesse de se développer – notamment grâce à son implication dans des projets internationaux tels que EHRI, qui réunit un consortium de 25 partenaires institutionnels (EHRI, 2020) – et qui pourraient potentiellement être activées pour faciliter la création de liens vers des ressources externes, ou encore pour favoriser l'enrichissement des données contenues dans l'instance Wikibase à l'aide de *données de la recherche* issues de partenaires.

Partenaires académiques

Le CegeSoma entretient d'étroites relations avec plusieurs départements d'histoire à travers toute la Belgique : cela laisse envisager des

1. À l'instar de la plateforme Belgium WWII (https://www.belgiumwwii.be/) ou de la plateforme The Belgian War Press (https://warpress.cegesoma.be/fr).

2. Comme en témoignent par exemple ces deux récentes publications comptant parmi leurs auteurs des chercheurs du Centre : Aerts *et al.* (2017); Maerten (2020).

3. Une note de vision – diffusée en interne – sur la politique d'ouverture des données des Archives de l'État explique ainsi que « moyennant un certain nombre de précautions, telles que des licences CC BY obligeant le réutilisateur à citer les AGR [Archives Générales du Royaume] en tant que source des données [...] une politique d'ouverture renforcera la visibilité en ligne de l'institution et de ses collections et donnera des Archives de l'État une image positive et progressiste. » (Depoortere et De Schamphelaere, 2019, p. 33). Cette politique consiste par exemple à proposer une réutilisation gratuite sous licence CC-BY d'une partie des métadonnées produites par l'institution, à l'instar des inventaires en XML/EAD (Depoortere et De Schamphelaere, 2019, p. 39).

enrichissements possibles de la Wikibase par le biais de séminaires dédiés ou encore de versement de jeux de données issus de projets de recherche basés sur les collections du Centre.

6.2 Faiblesses

Ressources limitées

Comme évoqué au cours de la présentation du contexte de cette étude de cas, les ressources humaines, financières et techniques de l'institution sont de plus en plus limitées [4], ce qui pourrait compliquer la maintenance d'une telle infrastructure.

Résistance au changement

Les modes de publications très ouverts et transparents, propres à un écosystème comme Wikibase, sont en rupture avec les formes de publications traditionnelles, qui se caractérisent par leur caractère très cadenassé. Ces particularités suscitent une certaine résistance au changement, encore accentuée par l'usage d'identifiants numériques abstraits et peu intuitifs, ainsi que par l'éventuel recours à des données issues de sources externes comme Wikidata.

Données en silos

Jusque-là, les données du CegeSoma n'ont pas du tout été reliées à des ressources externes. Cela signifie que d'importants efforts d'alignement de données sont encore requis au niveau des jeux de données préexistants avant de pouvoir pleinement exploiter le potentiel des données ouvertes et liées.

Données non standardisées

Les données préexistantes souffrent d'un manque de standardisation [5] et doivent donc faire l'objet d'un important travail d'harmonisation avant de pouvoir être intégrées dans l'instance Wikibase.

Données sommaires

Pour beaucoup de noms de personnes, les données préexistantes sont très sommaires [6], ce qui constitue un important obstacle à leur désambiguïsation et représente une faible plus-value informationnelle pour les utilisateurs.

4. Le Centre ne bénéficie par exemple plus de la présence d'un informaticien *in situ* depuis plusieurs années et doit se reposer sur les services d'appui des Archives de l'État.

5. C'est le cas par exemple des noms de lieux ou de réseaux de résistance encodés sous forme libre plutôt qu'à l'aide de référentiels.

6. Par exemple dans le cas du thésaurus Pallas, seuls les prénoms, noms et années de naissance et de décès sont connus.

Données peu structurées

Toute une partie des données préexistantes ne sont pas structurées et requièrent dès lors un important travail de pré-traitement afin que les éléments puissent être regroupés selon leur nature [7].

Granularité de l'indexation

Comme évoqué au cours des chapitres précédents, le niveau de granularité auquel l'indexation a été réalisée par le passé au CegeSoma a fait les frais de la migration des données vers le système des Archives de l'État et de l'information a malheureusement été perdue au cours de ce processus [8].

6.3 Opportunités

Besoins des utilisateurs

Comme souligné au cours du chapitre 3, le CegeSoma reçoit de plus en de plus de questions de particuliers se renseignant sur le passé de proches impliqués dans des activités de résistance ou de collaboration au cours de la Seconde Guerre mondiale, tandis que d'autres demandes plus ciblées proviennent de chercheurs travaillant par exemple sur les personnes liées à une zone géographique en particulier. Ces demandes s'inscrivent dans une tendance plus globale : les utilisateurs souhaitent désormais pouvoir accéder librement aux données *brutes*. Cet intérêt grandissant apparaît comme une opportunité à saisir pour optimiser l'accès aux données tout en rencontrant l'intérêt des utilisateurs.

Nombreux atouts du logiciel

Wikibase est un outil possédant de multiples avantages. Sans viser ici l'exhaustivité, soulignons le fait qu'il s'agit d'un logiciel libre et open source, qui possède une communauté active d'utilisateurs formant un réseau informel d'entraide ; qu'il offre beaucoup de flexibilité au niveau du modèle de données, permet de spécifier les références d'une information au niveau de granularité le plus fin ; qu'il est doté de fonctionnalités très poussées au niveau du contrôle des données ; et enfin qu'il est équipé de divers outils de visualisation de données. Autant d'atouts qui pourraient être utiles au CegeSoma.

7. Par exemple lorsque des alias et nom de naissance sont repris pêle-mêle, dans un même champ de texte, à la suite des noms, prénoms et années de naissance, ou encore dans le cadre des pages Personnalités de la plateforme Belgium WWII.

8. Cela signifie par exemple que des groupes de photographies auparavant distincts ont été rassemblés dans un même bloc et que seul le lien vers cette nouvelle cote est aujourd'hui disponible.

Politique Open Data

Le mouvement d'ouverture des données, qui se traduit tant par des initiatives propres au secteur du patrimoine culturel [9] que par des actes législatifs européens [10] impactant également les archives, apparaît comme une opportunité à saisir pour le CegeSoma pour s'affirmer dans une démarche de libre mise à disposition de ses (méta)données.

Positionnement stratégique

Si Wikibase est déjà répandu dans le milieu des bibliothèques – comme nous l'avons vu au chapitre 2 –, ce n'est pas encore le cas dans le secteur des archives. Il y a donc là une opportunité à saisir pour le CegeSoma qui pourrait ainsi renforcer le positionnement stratégique qu'il souhaite occuper au sein des Archives de l'État, mais également dans le paysage des institutions possédant des données liées à la Seconde Guerre mondiale, en s'affichant comme une sorte de *Linked Data Lab*.

Projet de norme *Records in Contexts*

Records in Contexts, la nouvelle norme de description archivistique en cours d'élaboration par l'ICA et déjà présentée en première partie de thèse, représente une opportunité de modernisation des données d'autorité. En effet, ce nouveau modèle de description repose sur l'utilisation d'entités désignées par des URIs. Les efforts déployés dans le cadre d'une Wikibase pourraient donc être valorisés dans la mise en application de cette norme.

Jeux de données externes

Les jeux de données mis en ligne dans des formats ouverts et pérennes par d'autres institutions représentent une opportunité d'enrichissement et de mise en contexte des données du CegeSoma, par le biais de la création de liens d'équivalence vers des ressources externes ou par la réutilisation des données préexistantes, telles les données Wikidata. Des liens d'équivalence peuvent également partir de ces ressources externes pour pointer vers les entités du CegeSoma, augmentant ainsi leur visibilité.

Synergies potentielles

Des synergies sont possibles à différentes échelles, que ce soit avec les partenaires d'autres projets développés par les Archives de l'État [11] ;

9. Comme *Open GLAM* : https://meta.wikimedia.org/wiki/Open_GLAM.

10. À l'instar de la troisème directive européenne sur l'Open Data du 20 juin 2019 (Journal officiel de l'Union européenne, 2019).

11. Comme la plateforme Belgium WWII, le projet Temas – pour *Thesaurus of Early Modern Archival Sources* –, la plateforme The Belgian War Press ou encore le projet Cartesius ; voir les pages web respectives de ces projets : https://www.belgiumwwii.be/ ; http://www.arch.be/index.php?l=fr&m=nos-projets&r=projets-de-recherche&pr=projet-temas-thesaurus-des-sources-d-archives-modernes ; https://warpress.cegesoma.be/fr ; http://www.cartesius.be/.

avec d'autres organismes belges ayant montré un intérêt pour Wikibase, à l'instar de Meemoo [12] ; avec des institutions partenaires également spécialisées dans l'histoire de la Seconde Guerre mondiale ; avec la communauté Wikibase ou encore avec la communauté Wikimedia Belgium.

Expertise technique des AGR

La récente intégration du CegeSoma aux Archives de l'État pourrait constituer une opportunité pour l'institution de bénéficier d'un soutien technique dans le cadre de la maintenance du serveur, mais peut-être également dans la mise en place de processus de synchronisation des données.

Inclusion d'autres types de données

L'une des plus grandes opportunités à saisir est certainement que le processus soit déclinable à d'autres entités. D'autres données pourraient être publiées et gérées via une telle base de connaissance, qu'il s'agisse d'entités personnes issues d'autres époques, de collectivités, de lieux ou encore de concepts.

6.4 Menaces

Instabilité du logiciel

Bien que le logiciel fasse actuellement l'objet d'une amélioration continue, il faut relever, outre les bugs actuels [13], que rien ne garantit – pour l'instant du moins – sa pérennité et sa gratuité au cours des années à venir. Or, comme Ribes et Finhold le soulignent, de tels systèmes doivent être pensés pour *the long now* (Ribes et Finholt, 2009, cité par Mattern, 2018).

Courbe d'apprentissage

La courbe d'apprentissage pour s'approprier la logique du modèle Wikibase pourrait menacer le succès d'une telle entreprise, d'autant plus que toute une partie des personnes susceptibles d'encoder des données dans cette base de connaissance sont des bénévoles majoritairement âgés. Il en va de même pour la prise en main des outils destinés à une importation massive de données, dont la complexité varie en fonction du type de données concernées [14].

12. The Flemish Institute for Archives : https://meemoo.be/en.

13. Voir par exemple les rapports de bugs publiés sur la plateforme Phabricator sous le tag *Wikibase-Containers* : https://phabricator.wikimedia.org/project/view/3079/.

14. Par exemple, les valeurs textuelles encodées dans l'interface QuickStatement doivent être encadrées par un nombre variable de guillemets en fonction de leur type, voir : https://www.wikidata.org/wiki/Help:QuickStatements/fr#Virgules_et_guillemets.

Sources externes de qualité variable

Dans le cadre de l'alignement des données vers des ressources externes, l'intégralité de ces ressources ne peut pas être contrôlée manuellement. Or, certains liens d'équivalence renvoient potentiellement vers des données incorrectes, ce qui pourrait se révéler d'autant plus problématique si ces valeurs inexactes sont récupérées dans le cadre de requêtes fédérées. En effet, elles viendraient ainsi menacer la qualité des résultats affichés, mais aussi potentiellement la crédibilité de la base de connaissance dans son ensemble, par exemple si la distinction entre les données produites en interne et les données récupérées n'est pas suffisamment communiquée aux utilisateurs finaux.

Maintenance de l'infrastructure

Outre la gestion des données et le contrôle de leur qualité, il faut également se soucier de la maintenance de l'infrastructure elle-même, sans quoi cette dernière sera susceptible d'être rapidement *cassée* [15]. En l'occurence, Wikibase repose sur le logiciel MediaWiki qui se caractérise par un modèle de développement *d'intégration continue* nécessitant des mises à jour régulières, les versions les plus anciennes n'étant pas maintenues indéfinement et ne bénéficiant par exemple plus de mises à jour de sécurité lorsqu'elles ont atteint le statut de *fin de vie* (MediaWiki, 2020a) [16].

Multiples données à synchroniser

Comme évoqué au cours des pages précédentes, si l'institution choisit de se reposer sur des ressources externes pour certains types de données ne constituant pas son cœur de métier, ces données seront rapidement obsolètes si elles sont seulement copiées dans la base de connaissance. Il est plus stratégique de mettre en place un système de synchronisation. Or, plus le nombre de sources à synchroniser sera élevé, plus la maintenance de ce système sera susceptible d'être sujette à des défaillances requérant une attention particulière.

Intégration aux workflows actuels

Le contexte de transition propre à la récente intégration du CegeSoma aux Archives de l'État s'accompagne de questionnements sur la façon dont les données contenues dans la Wikibase pourraient être intégrées aux workflows actuels, notamment dans le cas de producteurs d'archives, ces derniers devant être également encodés dans le

15. Comme le souligne Mattern : « Just like buildings and cities, most software applications and platforms and portals would break down quickly were it not for the maintenance workers who keep them in good working order. » (Mattern, 2018).

16. En l'occurence, « de nouvelles versions majeures sont publiées tous les six mois et les branches de ces versions reçoivent des mises à jour de sécurité jusqu'à un an après la première publication » (MediaWiki, 2020a).

logiciel de gestion utilisé par les Archives de l'État. Par ailleurs, il n'est pas encore clair de quelle façon les données de la Wikibase pourront être intégrées au sein du moteur de recherche des Archives de l'État étant donné que ce dernier fait actuellement l'objet d'une refonte.

Performance et scalabilité

Des problèmes de performance ayant déjà pu être constatés lors du lancement de requêtes SPARQL fédérées portant sur un nombre limité d'entités [17], des difficultés plus conséquentes sont donc susceptibles de menacer le bon fonctionnement du service de requête si l'instance est utilisée simultanément par un nombre plus conséquent d'utilisateurs et qu'elle accueille par ailleurs un nombre croissant de données.

Pérennité des URIs

Une Wikibase n'a que peu de sens si les entités identifiées à l'aide d'un URI ne sont pas pérennes . Or, si la base de connaissance passait du statut de prototype né dans le cadre d'un projet de recherche [18], à un statut plus pérenne et officiel, l'URL changerait, mais elle pourrait également changer en passant d'un projet estampillé *CegeSoma* à un projet géré par les *Archives de l'État*. Si un tel changement peut sembler anecdotique – quelques lettres d'un nom de domaine à changer –, il serait contraire aux bonnes pratiques (Berners-Lee, 1998), aurait un impact sur des milliers de données et menacerait la pérennité des URIs et par là même le sens de cette Wikibase.

6.5 Recommandations

Cette analyse SWOT vise à esquisser comment l'institution pourrait aligner ses activités internes sur les réalités externes (Gürel et Tat, 2017). La prochaine section vise donc à émettre des recommandations basées sur les forces, faiblesses, opportunités et menaces décrites au cours des paragraphes précédents. Bien qu'une partie de ces recommandations soient spécifiques au cas étudié, elles incluent toutefois des considérations généralisables à d'autres cas, dans la mesure où, comme expliqué dans le chapitre dédié à notre méthode (p. 31) :

> Nous pouvons dès lors considérer que ce qui fonctionne pour un petit centre de recherche et de documentation tel que celui-ci pourrait être généralisé à des cas similaires, mais également

17. Alors pourtant que l'installation repose sur 8GB de mémoire RAM, comme recommandé.

18. Avec une URL – que nous n'avons pu choisir – directement inspirée du nom de ce projet : adochs.arch.be.

> à des cas bénéficiant d'infrastructures et de moyens plus conséquents [19].

Les recommandations issues des observations menées dans le cadre de cette étude de cas et de l'analyse SWOT sont proposées sous la forme d'une stratégie par paliers. En effet, proposer des mesures graduelles semble opportun pour tenir compte des faiblesses et menaces énoncées au cours des sections précédentes, tout en capitalisant sur les forces et opportunités en présence. Cela permet d'envisager des aménagements progressifs en ce qui concerne la gestion des données d'autorité, au-delà de la question plus binaire et immédiate de l'adoption d'une Wikibase ou non. Les prochains paragraphes présentent les recommandations associées à chacun de ces trois paliers.

Le premier palier englobe les mesures les plus urgentes à mettre en place pour résorber la dette sémantique et éviter qu'elle ne continue à s'accumuler, dans une perspective où les moyens actuels ne permettraient pas, à court terme, d'implémenter une instance Wikibase. La recommandation principale est de commencer modestement en adoptant une optique de *good enough* [20], c'est-à-dire en se concentrant sur les briques nécessaires à l'élaboration de projets plus ambitieux – comme l'implémentation de la norme de descriptions archivistiques Records in Contexts ou la mise en place de nouveaux services aux usagers – : des métadonnées de qualité. Il s'agit donc d'agir, d'une part, sur les données préexistantes, et, d'autre part, à la source, c'est-à-dire sur les futures données devant être encodées.

Au niveau des données préexistantes, les efforts devraient se concentrer sur le travail amorcé dans le cadre de cette thèse, à savoir les différentes tâches décrites au cours du chapitre 4 et englobant le nettoyage, la structuration, la standardisation, le dédoublonnage et la réconciliation des données actuelles. Ces tâches de déduplication et de réconciliation concernent en priorité les noms de personnes, qui ne sont encore à ce stade que des chaînes de caractères qu'il s'agit de dédoublonner à l'aide de méthodes d'*entity linking* et de désambiguïser à l'aide d'URIs issus d'entités externes comme Wikidata ou VIAF [21]. Il faut toutefois être lucide sur le fait que de tels efforts sont plus ou moins conséquents selon la taille du jeu de données en présence, le degré de notoriété des personnes concernées et la richesse des données additionnelles pouvant être utilisées pour faciliter le processus de désambiguïsation. Par ailleurs, ce processus de sémantisation peut s'appliquer à d'autres données, stockées sous forme de chaînes de caractères dans diverses

19. À l'exception peut-être de questions liées plus spécifiquement à d'importants volumes de données, qui pourraient nécessiter des vérifications ultérieures.

20. Telle que décrite dans notre cadre théorique.

21. À terme, l'objectif serait évidemment de pouvoir se servir directement des URIs associés aux entités Personne du CegeSoma.

langues et qui gagneraient à être enrichies par des URIs équivalents. C'est le cas par exemple des noms de lieux ou des noms de professions. Dans de tels cas, l'usage d'un référentiel spécialisé pourrait potentiellement être jugé préférable à l'utilisation d'une base de connaissance généraliste telle que Wikidata, comme nous l'avons évoqué au cours de la section 4.3.1, mais cela nécessite toutefois de commencer par effectuer une analyse des ressources disponibles afin d'identifier en collaboration avec des experts du domaine ce qui répond le mieux aux besoins de l'institution.

Plus généralement, il faut accepter qu'il s'agit d'un équilibre à trouver lors de ce processus de *rétroconversion* des métadonnées : un arbitrage de type coûts-bénéfices doit être mené en tenant compte du volume du jeu de données à traiter, du nombre de champs disponibles, de la nature des informations disponibles, du temps de traitement requis, et du retour sur investissement pouvant être envisagé, comme par exemple l'exploitation qui pourrait être faite de ces données si elles étaient rendues disponibles dans un format plus facilement exploitable par des machines. Il faut également rester critique et garder à l'esprit que si l'augmentation du nombre de données exploitables permet par exemple d'aller plus loin dans l'analyse de questions statistiques, elle se traduit également par une complexité croissante des filtres ou outils de recherches.

Il serait par ailleurs contre-productif d'attendre que cette étape soit terminée avant de s'atteler à la suite, dans la mesure où la *dette* continuerait à se former. Nous préconisons plutôt d'envisager cela comme une tâche destinée à être poursuivie de façon constante, en toile de fond, de la même manière que des milliers d'inventaires d'archives et autres instruments de recherche font l'objet d'une *rétroconversion* afin que leur format puisse être adapté aux standards XML [22].

En ce qui concerne les nouvelles données, l'institution peut prévenir la formation de dette sémantique en traitant les problèmes à la source, c'est-à-dire au moment de la création de nouvelles données. Il s'agit de tenir compte du workflow actuel et d'identifier les premières mesures qui pourraient être mises en place. Au moment de la rédaction de cette thèse, ce sont principalement des bénévoles du CegeSoma – pour la plupart âgés – qui encodent de nouvelles données sur des personnes physiques, et plus précisément sur des personnes impliquées dans des activités de résistance au cours de la Seconde guerre mondiale. Pour ce faire, ils remplissent des tableaux Excel, qui sont ensuite considérés comme des annexes aux inventaires des fonds d'archives correspondants. En concertation avec les membres de l'équipe scientifique

22. Un travail de fond qui s'étend sur des années, mais qui a toutefois connu une vive accélération lors de la période de confinement qu'a connu la Belgique au cours du printemps 2020 – en raison de la situation sanitaire –, comme l'expliquent les Archives de l'État qui ont pu mettre en ligne plus de 1 000 inventaires rétroconvertis (Archives de l'État, 2020).

du Centre, il a été jugé plus stratégique de commencer dans un premier temps par maintenir ce fonctionnement [23], en veillant toutefois à standardiser les pratiques actuelles pour favoriser leur lecture par des machines. Certaines mesures ont commencé à être implémentées au CegeSoma au cours de l'été 2020 [24].

L'application de ces recommandations permet d'atteindre le premier palier. Il faut toutefois garder en tête qu'il s'agit seulement d'un palier intermédiaire et que les résultats sont dès lors limités. Ce travail de standardisation des données ne règle par exemple pas la question de leur publication : pour l'instant les jeux de données sont plus structurés, potentiellement enrichis, un peu plus documentés – à l'aide de métadonnées dédiées –, mais toujours constitués de fichiers épars. Comme nous le verrons, le deuxième palier vise à les importer dans une Wikibase et ainsi les doter d'un URI. Cependant, si cette stratégie visant à intégrer ces données à une Wikibase devait être jugée inenvisageable, deux pistes alternatives se présenteraient pour quand même aller plus loin au niveau de la publication et surtout de l'interconnexion de ces données. La première piste consiste à publier les données sous forme de *Linked Data* sur le site même de l'institution, comme l'a par exemple illustré

23. À l'avenir, une autre piste à explorer pourrait être l'utilisation Google Spreadsheets intégrant des fonctionnalités de liens directs à Wikidata (voir Steiner, 2016).

24. À titre indicatif, voici une liste des premières mesures ayant été mises en place en ce qui concerne ces fichiers Excel :

- l'ajout d'un champ *remarques* afin d'éviter que ne soient mélangés des valeurs et des commentaires textuels au sein d'une même cellule
- la systématisation du fait que chaque cellule ne doit contenir qu'un seul type de données : de nouvelles colonnes sont dès lors créées en cas de besoin (par exemple dans le cas de surnoms ou pseudonymes qui étaient accolés jusque-là au prénom)
- l'encodage du prénom principal dans une colonne distincte des éventuels autres prénoms
- l'ajout de toute information d'identification pouvant être utile dans le cadre de la désambiguïsation, même si cela ne semble pas directement utile dans le contexte direct d'utilisation des données (par exemple lorsqu'une date de décès se situe bien après la fin de la guerre)
- la standardisation des champs de date et de langue afin qu'ils soient conformes aux normes ISO correspondantes
- la standardisation des noms de lieux en Belgique à l'aide du référentiel utilisé par les Archives de l'État
- la systématisation des abréviations utilisées dans le cadre de valeurs présentes en nombre restreint (par exemple pour identifier le genre d'une personne) et, dans les autres cas, la création de référentiels (par exemple dans le cas de réseaux et mouvements de résistance) pouvant être utilisés sous forme de listes déroulantes équipées d'une fonction d'auto-complétion, afin de limiter le nombre de champs contenant du texte libre
- la création d'une feuille Excel dédiée aux métadonnées portant sur le jeu de données lui-même, afin de spécifier le contexte de production des données, d'identifier les fonds d'archives associés aux personnes décrites et de spécifier les caractéristiques et mesures à respecter pour chaque colonne ; un aperçu est proposé dans l'Annexe 13, p. 402.

.

Meemoo – the Flemish Institute for Archives – dans le cadre de la publication de la presse numérisée de la Première guerre mondiale[25]. Une autre piste serait de publier une partie de ces données[26] sur Wikidata, à l'instar de ce qui a été fait dans le cadre du projet WeChanged. Le procédé a consisté a importer sur Wikidata par le biais d'une propriété spécifique – WeChangEd ID|P7947[27] – plus de 3 500 entités et relations issues de la base de données du projet, de manière à faciliter la réutilisation et l'intégration de ces informations avec d'autres sources de *Linked Open Data* (Birkholz, 2020).

Comme évoqué dans l'introduction, cette section est structurée autour de trois paliers de recommandations. Ce second palier propose des directives et pistes visant à faciliter l'utilisation d'une instance Wikibase pour la gestion quotidienne des données d'autorité.

Tout d'abord, il s'agira d'ajuster le prototype créé dans le cadre du projet d'Adochs, en commençant par le doter d'un nouveau nom de domaine qui gardera du sens même si le projet s'agrandit – de manière à ce que la pérennité des URIs soit garantie –, mais également en effectuant les diverses modifications nécessaires, comme l'éventuelle modification du type de données pouvant être acceptées comme valeurs d'une propriété. Pour favoriser l'adoption de l'outil par le personnel et pour composer avec une certaine résistance au changement, nous recommandons d'intégrer le personnel scientifique à ce type de réflexions – concernant la modélisation des données –, mais également de développer des exemples éloquents permettant de démontrer l'utilité de ces nouvelles pratiques. Ce dernier point est toutefois délicat dans la mesure où il faut convaincre et solliciter la participation du personnel sans avoir forcément encore de quoi convaincre, étant donné que la participation du personnel est parfois requise pour pouvoir avancer et atteindre des étapes concrètes[28].

25. Comme le soulignent les auteurs de ce projet : « To publish linked data, you can choose to use a separate interface (for example a SPARQL endpoint or REST API) or datadump or simply enrich your own website. We prefer the last option. It's easier to find the way when there's only one address. Moreover, a separate interface requires additional maintenance, which entails the risk that your website might be more up-to-date than your interface. Users or developers can easily, almost automatically, examine the website, scrape and access datasets by using JSON-LD for the publication. Data on the website is readable both for humans (the website on display when you visit newsfromthegreatwar.be) and machines. » (meemoo, 2019). Pour consulter les détails techniques, voir également : https://brechtvdv.github.io/Article-Using-an-existing-website-as-a-queryable-low-cost-LOD-publishing-interface/.

26. C'est-à-dire toutes celles qui répondent aux critères de notoriété de Wikidata.

27. Voir le formulaire de proposition de création de cette nouvelle propriété : https://www.wikidata.org/wiki/Wikidata:Property_proposal/WeChangEd_ID.

28. Ce phénomène touche par exemple également Wikidata dans la mesure où « the actual usefulness of the data comes with its use. [...] Without complete and high-quality data, there are no cool apps. But without interesting apps, there are few incentives to provide data and to improve its quality » (Estermann, 2018).

Au niveau du remplissage de l'instance Wikibase, l'essentiel est de commencer de façon modeste mais raisonnable, en travaillant par exemple sur un sous-ensemble, en ayant toutefois déjà à l'esprit une future expansion thématique ou chronologique : les propriétés doivent en effet être pensées dès le départ de la façon la plus globale possible.

En ce qui concerne la courbe d'apprentissage susceptible de menacer une bonne adoption de l'outil, des mesures peuvent être mises en place à deux niveaux afin de réduire la complexité. Premièrement, à la source, en améliorant l'interface utilisateur, que ce soit en adaptant des outils Wikidata [29] ou en travaillant sur l'utilisation de formulaires facilitant l'encodage des données [30]. Deuxièmement, il est possible d'agir au niveau de la transmission de connaissances en organisant des ateliers destinés aux personnes s'occupant de l'encodage, en mettant en place des tutoriels vidéos ou encore en envisageant la mise en place d'un services de requêtes (SPARQL) à la demande, à l'instar de ce que propose Wikidata [31].

Au niveau de la qualité de l'information et plus précisément de la qualité des alignements vers des identifiants externes, il est recommandé d'inclure des scores de réconciliation sous la forme de qualificatif Wikibase afin de pouvoir documenter le degré d'incertitude et éviter de *tromper l'utilisateur sur la marchandise*. De même, si des éléments sont décrits à l'aide d'informations issues de sources externes telles que Wikidata, il est alors primordial de préciser la provenance de ces informations, ce qui permet à l'institution d'assurer une transparence quant à l'origine de l'information, tout en se dédouanant des éventuels problèmes de qualité de ces données – en distinguant clairement les informations issues de l'institution de celles issues de l'extérieur.

Plus globalement, les droits d'accès et de modification de l'instance Wikibase doivent être finement configurés de manière à pouvoir assurer le suivi et le contrôle des nouvelles informations encodées, mais également du modèle de données – c'est-à-dire les propriétés et qualificatifs utilisés pour décrire les éléments –, des informations annexes comme la documentation et enfin des gadgets et extensions visant à améliorer l'expérience utilisateur.

Par ailleurs, il semble primordial de nommer une personne responsable du suivi et de la qualité des données d'autorités, mais également de pouvoir insérer ces activités de gestion et de contrôle dans le cahier des charges de

29. Comme Description, Recoin ou QuickPresets, voir également cette liste : https://www.wikidata.org/wiki/Wikidata:Tools/Enhance_user_interface.

30. Comme le permet par exemple Cradle, qui est cependant encore en développement et peu convaincant pour l'instant, voir : https://www.wikidata.org/wiki/Wikidata:Cradle.

31. À travers sa page intitulée Request a query, voir : https://www.wikidata.org/wiki/Wikidata:Request_a_query.

tous les membres du personnel concerné, afin que cette tâche ne soit pas invisibilisée et puisse être intégrée aux tâches structurelles [32].

Au niveau de la réutilisation des données, notamment dans le cadre des infrastructures préexistantes des Archives de l'État, une première étape – destinée à être poussée plus loin dans le cadre du troisième palier – pourrait consister à prévoir des requêtes SPARQL permettant d'exporter les données désirées en JSON, en XML ou en CSV. Alternativement, comme illustré au cours de la sous-section 5.2.2, un script peut être utilisé afin de générer des notices d'autorité en XML au format EAC-CPF à partir des données contenues dans l'instance Wikibase.

En ce qui concerne la maintenance de l'infrastructure et les questions de performance, il pourrait être intéressant pour l'institution de se mettre en relation avec la Fondation Wikimedia afin de mettre en place un partenariat officiel et d'éventuellement pouvoir bénéficier d'un soutien technique [33]. Une autre piste serait d'envisager l'utilisation de *Wikibase as a service* [34] afin de se décharger des tâches les plus techniques, sous réserve de l'évolution du service [35].

Enfin, il est recommandé de mener une veille afin de pouvoir suivre le développement du logiciel, que cela soit en restant au contact de la communauté d'utilisateurs Wikibase – qui constitue un véritable réseau d'entraide et de partage d'expérience, comme évoqué au cours de la section 2.1.2 –, en consultant la littérature scientifique à ce sujet ou en parcourant la feuille de route de l'équipe de développement de Wikidata et Wikibase.

Si ces recommandations permettent d'épauler l'utilisation quotidienne d'une instance Wikibase pour la gestion des données d'autorité, cette utilisation reste toutefois limitée. Elle est destinée à être perfectionnée, approfondie et déployée plus largement lors du passage au troisième palier. La priorité dans le cadre de ce troisième palier sera de systématiser et de fluidifier le workflow complet, de la création des données à leur mise à disposition des utilisateurs finaux.

En parallèle des possibilités de création manuelle de données à l'aide de l'interface graphique, il serait bénéfique de systématiser l'importation de

32. Comme l'ont souligné les 70 représentants du secteur culturel ayant participé au forum organisé dans le cadre du GND for Cultural Data (GND4C), la transformation numérique dans le secteur culturel doit être vue comme une tâche structurelle et des opérations comme la description du matériel numérisé à l'aide de métadonnées elles-mêmes liées à des données d'autorité nécessitent un financement permament et continu (Fischer et Manecke, 2019).

33. Ce dont a par exemple bénéficié la Bibliothèque nationale allemande : « Since then WMDE has had workshops and regular syncs with the German National Library, and will continue doing so, to support them in migrating to Wikibase. » (Pintscher *et al.*, 2019b, p. 12).

34. https://www.wbstack.com/.

35. Comme souligné à la section 2.2.1, les questions de pérennité et d'éventuels coûts – pour l'instant le service n'est pas payant – devraient être clarifiées au cours des mois à venir.

nouveaux jeux de données. Par exemple si les données sont préalablement encodées dans des fichiers CSV, il faudrait rationaliser la mise en correspondance de l'information avec les données issues de Wikibase [36]. Il faudrait également fluidifier le processus de détection de doublons et d'importation massive des données dans la Wikibase, ce qui pourrait se faire en passant par le logiciel de traitements de données OpenRefine [37], évitant ainsi de devoir passer encore par une autre interface telle que QuickStatements.

Notons que lorsque ce workflow de création des données aura été affiné et éprouvé, la voie pourra progressivement être ouverte à d'autres jeux de données, que ce soit en interne au niveau du CegeSoma, ou à l'échelle plus large des 19 dépôts que comptent les Archives de l'État en Belgique. Il sera en effet plus facile à ce stade de convaincre de l'intérêt d'une telle infrastructure, et l'inclusion d'autres types de données pourra être pensée en tirant parti des enseignements et de l'expérience liée à la gestion des entités personnes. Dans cette optique, il est recommandé d'établir des critères de priorisation des jeux de données à mettre en ligne sous forme de Linked Open Data, tels que l'utilisation potentielle de ces données, la faisabilité technique et juridique, ainsi que la volonté des détenteurs de données (Estermann *et al.*, 2020, p. 9-10).

En ce qui concerne le workflow global, il faut garder à l'esprit que les données importées dans la Wikibase devraient pouvoir être affichées au sein d'une interface de consultation plus conviviale, Wikibase étant une infrastructure avant tout destinée à pouvoir être interrogée par des machines. Cela pourrait se faire en réutilisant des outils inspirés de Wikidata, comme Reasonator [38] ou Sciencesstories.io [39], comme cela a notamment été fait dans le cadre du projet gantois WeChanged [40], ou en créant une interface *front end* dédiée, comme le proposent par exemple le projet LCA (Leibniz's Correspondents and Acquaintances) [41], le projet BERD@BW (Business and Economics Research Data Center Baden-Württemberg) [42], ou encore le pro-

36. Une première étape pourrait par exemple consister à utiliser une liste déroulante incluant le libellé et son identifiant numérique Wikibase ; mais il faudrait investiguer plus avant la façon dont ces données pourraient être automatiquement synchronisées et mises à jour en passant par exemple par l'API associée à la Wikibase.

37. Dont le système d'ajout de données dans Wikidata a été très récemment configuré afin de pouvoir également être relié à d'autres instances Wikibase, comme énoncé au cours du chapitre précédent.

38. Voir : https://reasonator.toolforge.org/.

39. Qui a par ailleurs obtenu le prix LODLAM (Linked Open Data in Libraries, Archives and Museums) en 2020. Voir : http://sciencestories.io/.

40. Voir : https://stories.wechanged.ugent.be/.

41. Voir : https://leibnitiana.eu/architecture.

42. Voir : https://madoc.bib.uni-mannheim.de/54900/1/2020-05-05_KIM_BERD_Shigapov.pdf.

jet KOHESIO (the project Information Portal for EU Cohesion Policy) [43]. Dans le cas précis des données du CegeSoma et de leur intégration au sein du futur moteur de recherche des Archives, notre recommandation serait d'utiliser les URIs Wikibase identifiant des personnes directement dans les inventaires EAD, afin de pouvoir ensuite les utiliser pour générer des requêtes SPARQL, qui renverraient l'information désirée – comme par exemple des dates de naissance et de décès – au format souhaité. Ce processus pourrait également inclure des requêtes fédérées permettant d'aller rechercher des données issues de Wikidata ou d'autres bases de connaissance, comme illustré à la sous-section 5.2.3.

Enfin, même si l'instance Wikibase est considérée comme l'espace principal de création et de gestion des données d'autorité [44], il n'en reste pas moins que certaines d'entre elles seront peut-être issues de sources externes. Comme suggéré au cours des chapitres précédents, il faudra alors mettre en place un système de synchronisation en s'inspirant des bonnes pratiques [45].

En parallèle du perfectionnement de ce workflow, d'autres actions pourront être entreprises. Tout d'abord, il serait certainement très utile de pouvoir implémenter l'extension Wikibase Quality Constraints [46], évoquée au cours du chapitre 5, afin de prévenir les problèmes de qualité des données et de pouvoir systématiser leur détection.

De plus, comme suggéré au cours des pages précédentes, il pourrait être intéressant de combiner l'utilisation de méthodes de *Named Entity Recognition* avec les possibilités offertes par le service de réconciliation basé sur l'instance Wikibase afin de pouvoir lier des contenus composés de texte non structuré – comme par exemple la presse clandestine datant de la Seconde mondiale et publiée en ligne par le CegeSoma dans le cadre du projet the Belgian War Press [47] aux entités Personne de la Wikibase.

En ce qui concerne l'opportunité de réutilisation des données stockées dans la Wikibase dans le cadre de la nouvelle norme archivistique Records in Contexts, cette étape pourrait être facilitée en poursuivant les efforts d'alignement entre les propriétés utilisées au sein de la Wikibase – principalement issues de Wikidata – et les propriétés utilisées par l'ontologie Records in Context, tels qu'entrepris à la sous-section section 4.3.4.

43. Voir : https://kohesio.eu/.

44. Et pas seulement a second home – c'est-à-dire un entrepôt destiné à accueillir une copie de données issues d'autres base de données –, comme le sont d'autres instances Wikibase, à l'instar de l'instance sur laquelle travaille la Bibliothèque nationale allemande (Fischer, 2018).

45. Voir par exemple les travaux réalisés par Andra Waagmeester dans le cadre du projet Gene Wiki pour synchroniser de multiples sources de données (Waagmeester, 2019, p. 25-28) ; (Waagmeester *et al.*, 2020).

46. https://github.com/wikimedia/mediawiki-extensions-WikibaseQualityConstraints.

47. https://warpress.cegesoma.be/fr.

D'autre part, l'institution pourrait gagner en visibilité en faisant profiter Wikidata des informations stockées dans son instance, que cela soit en intégrant directement ces données dans Wikidata [48], et, ou en soumettant à la communauté Wikidata la proposition de création d'une propriété permettant d'associer aux éléments Wikidata un identifiant externe propre au CegeSoma et, ou aux Archives de l'État.

Enfin, il est clair que l'institution aurait tout à gagner en mettant en place des synergies. Ces synergies sont possibles à différents niveaux : par exemple avec des partenaires institutionnels afin d'établir des liens mutuels entre des entités et ressources jusque-là isolées et ainsi offrir une meilleure contextualisation des collections. Cela peut également passer par des collaborations avec des partenaires académiques. Bien qu'il ne s'agisse pas d'une priorité en soi et que la fonction première de la Wikibase reste de pouvoir stocker et gérer des données d'autorité, une extension possible pourrait consister à intégrer des données de la recherche à même d'apporter des précisions sur les entités déjà présentes dans la Wikibase. Une autre piste prometteuse concerne le potentiel éducatif d'une telle base de connaissance : elle offre un support concret pour initier des étudiants au potentiel des données structurées à l'aide d'exemples tirant parti du service de requêtes SPARQL [49]. Cela permettrait à l'institution d'élargir ses activités d'histoire publique et de valorisation des collections tout en bénéficiant des analysées réalisées dans ce cadre [50]. Dans la continuité de cette idée, il pourrait également être bénéfique de mettre en place de façon régulière des éditathons, c'est-à-dire des sessions d'édition collaborative des entités présentes dans la Wikibase afin d'améliorer et d'enrichir les données. Enfin, une dernière recommandation serait de prendre part activement aux réseaux de partage de connaissance liés à l'écosystème Wikidata-Wikibase, à l'instar des réunions bimensuelles organisées par le LD4 Wikidata Affinity Group Call [51], voire même de contribuer à leur création ou mise en œuvre dans un contexte européen [52].

48. S'il est évidemment possible de le faire *manuellement*, il pourrait être stratégique d'attendre que les efforts de fédération entre instances Wikibase aient progressé afin d'éviter les défis de maintenance et de synchronisation évoqués au cours des pages précédentes.

49. Comme le souligne Martin Poulter, le langage de requête SPARQL peut être enseigné comme une langue étrangère. Il a notamment montré comment le service de requête de Wikidata pouvait être utilisé pour enseigner ce type de représentation de l'information à des publics non techniques (Poulter, 2019).

50. Voir par exemple les productions réalisées dans le cadre d'initiatives telles que *Wikidata in the Classroom* : https://blog.wikimedia.org.uk/2018/03/data-on-the-history-of-scottish-witch-trials-added-to-wikidata/.

51. https://wiki.lyrasis.org/display/LD4P2/LD4-Wikidata+Affinity+Group.

52. Par exemple dans la continuité des webinaires organisés au cours du printemps 2020 par l'Association des archivistes français·e·s (AAF), dans le cadre de son partenariat avec Wikimédia France, voir : https://meta.wikimedia.org/wiki/WikiArchives. La session du 14 mai 2020 était ainsi dédiée à Wikidata et Wikibase : https://meta.wikimedia.org/wiki/WikiArchives/Idées_Wikibase_pour_archivistes.

Conclusions et perspectives

Au départ de cette thèse, il y avait cette interrogation : dans quelle mesure les données d'autorité et les vocabulaires contrôlés ont-ils toujours leur sens dans le contexte du Web de données, et, le cas échéant, comment sont-ils appelés à évoluer ? Ce questionnement initial s'est ensuite affiné au contact de la *maintenance theory*, présentée en introduction, pour aboutir à cette question de recherche :

> *Comment favoriser une gestion soutenable des données d'autorité archivistiques dans le cadre du Web de données ?*

Pour aborder cette question de recherche avec nuance, il était nécessaire de pouvoir adopter une position permettant à la fois une prise de distance critique et une forte proximité avec le sujet. Cela fut permis grâce à une approche combinant à la fois l'analyse d'un état de l'art approfondi et une étude de cas basée sur des données empiriques. En effet, notre question de recherche visant à répondre à la question du *comment*, il était dès lors indispensable de pouvoir éprouver les pistes rencontrées dans le cadre de la première partie avec des données issues du monde réel, qui plus est dans le cadre d'une étude de cas à même de susciter des recommandations généralisables. C'est ce que nous avons cherché à faire, en confrontant les belles promesses du Web sémantique à une réalité de terrain qui s'en trouve parfois extrêmement éloignée.

En effet, le travail sur les données est loin d'être toujours propre : « when it comes to data, there really is no way around getting your hands dirty » (Tapley Hoyt, 2020). Cela implique du travail *manuel*, de nombreux ajustements et compromis. Or, à nos yeux, c'est précisément cette immersion du scientifique dans la réalisation de tâches laborieuses et répétitives et sa confrontation à des questions d'apparence très prosaïques, qui sont à même de susciter de sa part une compréhension plus fine de l'importance du travail de maintenance. Ce travail, qui se caractérise par son invisibilité et ses effets difficiles à se représenter [53], est en effet susceptible d'être sous-estimé s'il n'a pu être expérimenté de l'intérieur.

Dans le cadre de cette thèse, ce travail s'est traduit par des étapes de manipulation et de traitement de données destinées à être modélisées et importées dans une base de connaissance reposant sur le logiciel Wikibase. Ces étapes, détaillées au cours des chapitres 3, 4 et 5, ont fait l'objet d'une analyse approfondie et de recommandations détaillées dans le cadre du chapitre 6. Elles nous permettent également d'apporter des éléments de réponse

53. Comme le relèvent Dagiral et Peerbaye, « les situations de travail que connaissent les organisations distribuées aux prises avec les infrastructures informationnelles contemporaines placent de plus en plus souvent les personnes dans des situations d'invisibilité du travail pour soi et pour autrui. L'opacité des outils techniques, le caractère hautement abstrait des micro-opérations se combinent avec leur nature répétitive, distribuée, d'apparence banale et aux effets difficiles à se représenter. » (Dagiral et Peerbaye, 2012, p. 212).

aux trois sous-questions de recherche présentées dans l'introduction de cette thèse, que nous reprenons ici.

> *QR1 : Dans quelle mesure est-il possible de se reposer sur des processus d'automatisation pour réduire la dette sémantique pesant sur les données d'autorité ?*

Cette première question visait à étudier les stratégies permettant d'agir sur la dette sémantique [54] plutôt que de la laisser s'accumuler.

Premièrement, il faut être lucide sur le fait que plusieurs actions sont souvent requises avant d'arriver à cette étape visant à lever l'ambiguïté sur le sens d'un mot. En effet, pour peu que plusieurs informations exprimées en langage naturel soient mélangées au sein d'une même cellule, il faut d'abord commencer par les réorganiser de manière plus structurée, ce qui requiert une analyse préalable ainsi que certains traitements manuels en cas de données peu standardisées.

Deuxièmement, il s'avère en effet possible d'automatiser ce processus de sémantisation en déléguant à un logiciel cette tâche. Il se chargera alors d'établir des scores de similarité entre des chaînes de caractères sujettes à l'ambiguïté et des données de référence – que ce soit des référentiels propres à l'institution ou des bases de connaissance externes comme Wikidata. Si cela permet un gain de temps conséquent, nous avons toutefois observé une tendance rappelant le principe de Pareto, qui veut que 80% des effets soient le produit de 20% des causes. Dans le cas précis, il s'agit d'une minorité de données auxquelles sont associés des scores incertains, qui représenteraient 80% du temps de traitement, dans la mesure où les vérifications manuelles sont extrêmement chronophages. Ainsi, le résistant belge *Richard Altenhoff* présent dans un de nos jeux de données possédait par exemple une date de naissance ne coïncidant pas avec celle de Wikidata [55], le score de réconciliation fut donc mitigé, rendant nécessaire une recherche approfondie et la consultation de sources externes pour confirmer ou infirmer ce lien d'équivalence.

Dans ce genre de cas, si les ressources manquent pour pouvoir effectuer des vérifications manuelles approfondies ou que leur volume est trop important, des solutions alternatives existent. En adoptant une approche de type *good enough*, il est en effet envisageable de fournir à l'utilisateur le plus d'informations possibles, sans passer sous silence l'ambiguïté ou l'incertitude entourant certaines données, ni mettre en péril la crédibilité de l'institution. Comme nous l'avons souligné au cours du chapitre 4, il s'agit de pouvoir indiquer de façon transparente que les données sont imparfaites

54. Notion introduite dans l'introduction, page 9.

55. La page Wikidata de l'élément Q3430462|Richard Altenhoff annonce une naissance en 1920, tandis que la plateforme Belgium WWII indique une naissance en 1913.

et qu'une certaine incertitude subsiste (Garmendia, 2019). Dans le cadre de notre instance Wikibase, nous avons déjà déployé un dispositif minimal [56], qui mériterait d'être approfondi afin d'éventuellement inclure directement le score issu du processus de réconciliation et, ou une typologie plus fine des différents types d'incertitude pouvant se présenter [57]. Cependant, il faut garder à l'esprit que plus l'information est précise, plus la complexité croît, que cela soit au moment de l'encodage [58] des données ou au moment de leur interrogation.

Troisièmement, il faut garder à l'esprit que la sémantisation des données à partir de référentiels pose la question de la source des données à utiliser, qui est loin d'être évidente – comme nous l'avons illustré au cours du chapitre 4 avec le cas des lieux et des métiers. Aux côtés de certaines étapes pouvant être automatisées subsistent donc des questionnements ne pouvant être esquivés et nécessitant analyses et arbitrages de type coût-bénéfice. Par ailleurs, comme nous l'avons exposé au cours des quatrième et cinquième chapitres, cela soulève également la question de la maintenance des données, notamment si certains référentiels sont destinés à être intégrés et gérés directement dans la Wikibase.

Quatrièmement, il faut garder à l'esprit qu'il ne s'agit pas d'une question exclusivement technique. Dans le cadre d'un traitement de la dette sémantique impliquant un passage à une logique de graphe d'entités, il s'agit d'un véritable changement dans la manière de modéliser l'information, qui repose désormais sur une structure en graphe. Les profils les plus techniques et les experts du domaine se retrouvent alors à entamer un dialogue afin de choisir ce qu'ils veulent exprimer, de quelle manière, et à ensuite mettre en place des procédures pour garantir la transparence de ces choix. Si la sémantisation des données apparaît dès lors comme un véritable vecteur de changement au sein de l'institution, il faut toutefois garder à l'esprit que cela implique des modifications des méthodes de travail pouvant susciter des résistances, d'autant plus que la récolte des bénéfices doit surtout être envisagée sur le long terme. Ces considérations soulèvent la question des outils disponibles pour faciliter ce travail, ce qui nous ramène à la deuxième sous-question de recherche de cette thèse :

56. La propriété P25|qualité de l'information (https://adochs.arch.be/wiki/Property:P25) peut être utilisée comme qualificatif, en étant par exemple associée à la valeur Q24774|probablement (https://adochs.arch.be/wiki/Item:Q24774).

57. À titre indicatif, l'instance Wikibase FactGrid possède par exemple une propriété – intitulée « how sure is this ? » – « to state the solidity of an argument », à laquelle sont associées pas moins de 14 valeurs possibles, voir : https://database.factgrid.de/wiki/Property:P155.

58. Dans le cadre de l'encodage de données dans des fichiers tabulaires, cela signifierait passer de ce qui est souvent une seule colonne de remarques en vrac, à une information beaucoup plus fine rattachée à chaque valeur, alourdissant considérablement le volume du document.

> *QR2 : De quelle manière les fonctionnalités offertes par le logiciel Wikibase peuvent-elles être utilisées pour faciliter et rationaliser le travail de création et de maintenance des données d'autorité ?*

Cette seconde question avait pour ambition de tester les capacités de Wikibase à faciliter les activités de maintenance des données. Les observations menées dans le cadre de la création d'un prototype Wikibase nous permettent de poser différents constats.

Comme l'ont souligné Lovins et Hillmann (2017) dans le cadre de la maintenance de vocabulaires bibliographiques, il est utile que les fonctions de maintenance soient directement intégrées aux outils et workflows quotidiens. Le logiciel Wikibase permet cela, en offrant des outils comme le *versioning*, qui permet à tout moment de consulter l'historique des opérations ayant été réalisées et d'éventuellement revenir à une version antérieure, ou encore les liste de suivi qui facilitent le contrôle de qualité des données, comme nous l'avons illustré au cours du cinquième chapitre [59]. Cela représente une rupture radicale avec des pratiques actuelles plus informelles, incluant énormément de prises de décisions discutées oralement ou dans le cadre d'échanges privés, qui ne peuvent dès lors pas être archivées et consultées par d'autres à l'avenir [60]. En sauvegardant toutes les opérations effectuées à un fin niveau de granularité, en les rendant publiquement accessibles par défaut et en proposant des outils de discussion directement rattachés aux objets concernés, un outil comme Wikibase a dès lors le potentiel d'augmenter significativement le *bus factor*. Ce concept, utilisé en développement logiciel pour mesurer les risques causés par un faible partage d'informations et de compétences enre les membres d'une équipe, désigne le nombre de personnes pouvant disparaître soudainement d'un projet – se faire renverser par un bus – avant qu'un projet n'échoue (Cosentino *et al.*, 2015).

Par ailleurs, l'enregistrement et l'affichage systématique d'opérations traditionnellement réalisées dans l'ombre permettent incontestablement de donner de la visibilité aux tâches de création et de maintenance des données. La possibilité de quantifier ces opérations représente ainsi une opportunité de valoriser des tâches d'ordinaire très invisibilisées et dès lors de disposer d'arguments en faveur d'une meilleure allocation de moyens structurels pour

59. L'exercice reste cependant avant tout théorique : les différents outils sont destinés à être testés avec les personnes concernées, ce qui n'a pu être fait faute de temps, ainsi qu'en raison du caractère éphémère du prototype – comme en témoignent les URIs basés sur un nom de domaine lié spécifiquement à ce projet de recherche, qui sont destinés à évoluer.

60. Cela signifie qu'il s'agit également d'accompagner le changement de mentalité qu'entraîne une publication en ligne en toute transparence, en reconnaissant qu'il peut s'agir pour le personnel d'un important investissement à concéder, sans que la plus-value ne soit forcément directement visible.

leur réalisation [61]. Cependant, il faut garder à l'esprit que la précision de cet historique permet également de suivre les activités de chacun des contributeurs et ces informations pourraient dès lors être utilisées à des fins de contrôle de la productivité des employés.

Enfin, si un logiciel comme Wikibase fournit tout un arsenal d'outils pouvant contribuer à un meilleur traitement de la qualité des données, nous attirons l'attention sur le fait que cela pourrait être source de dispersion ou de surchage informationnelle. La multiplication des sources d'information à consulter, contrôler voire même commenter, pourrait ainsi s'avérer dissuasive pour des personnes dont la gestion des données d'autorité ne constitue qu'une partie du travail. Il s'agit donc de tâtonner afin de déterminer ce qui fonctionne et est réellement utile dans un certain contexte, mais également d'envisager les possibilités de mutualisation, abordées dans le cadre de la troisième sous-question de recherche :

> *QR3 : Comment les Linked Open Data peuvent-elles faciliter de nouvelles formes de mutualisation susceptibles de réduire le volume de données à maintenir ?*

Cette troisième question de recherche visait à observer de façon très concrète les possibilités de partage ou de réutilisation de données offertes par les technologies des données ouvertes et liées, en 2020. Les expérimentations menées dans le cadre de notre étude de cas nous ont permis différentes observations. Tout d'abord, il s'avère que ce n'est pas encore évident dans les faits. Sachant que le projet de fédération d'instances Wikibase est encore en cours de développement, les efforts de synchronisation entre les données de différentes instances s'avèrent donc laborieux et constituent finalement une nouvelle forme de maintenance devant être assurée. Dans le cadre par exemple de la réutilisation de données Wikidata, il s'agit d'opérer un arbitrage de type coût-bénéfices afin de déterminer s'il est préférable d'importer ces données pour les utiliser sous forme de nouveaux éléments de la Wikibase ou plutôt de les utiliser sous forme d'identifiants externes.

De plus, si des données externes peuvent être récupérées à l'aide de requêtes fédérées, il faut garder à l'esprit que cela nécessite de disposer d'un pivot entre les deux sources concernées. Or cette étape d'alignement mobilise d'importantes ressources, comme nous l'avons souligné. Cela requiert de surcroît que les données soient librement accessibles par le biais d'un point d'accès SPARQL. Or, cette pratique n'est pas encore très largement répan-

61. Bien que l'accès numérique aux ressources et à leurs collections soient considérés comme une tâche essentielle des institutions du patrimoine culturel, il apparaît que le financement permettant la réalisation de ces tâches provient en effet souvent de sources non structurelles (Liew, 2009, cité par van Hooland *et al.*, 2011).

due dans le contexte des archives et du patrimoine culturel [62]. Sachant que l'intérêt des données liées croît avec le nombre de jeux de données partagé, cela signifie qu'en fonction du domaine concerné, les bénéfices pouvant être tirés de telles requêtes seront peut-être faibles ou inexistants.

Enfin, cela entraîne de nouveaux défis au niveau de la conception des interfaces de recherche, qui doivent pouvoir afficher de façon claire et transparente la provenance de données issues de sources externes, mais également s'adapter au fait que ce type d'information ne sera pas systématiquement disponible et que sa qualité varie.

Après avoir passé en revue chacune de ces trois sous-questions de recherche, nous pouvons maintenant revenir à la question principale et tenter d'y répondre à la lumière de ces constats et de considérations plus larges ayant pris naissance dans le cadre de cette recherche.

Mais avant d'apporter des éléments de réponse, commençons toutefois par souligner le fait que la tension opposant les problèmes de maintenance à l'innovation, telle que décrite dans notre cadre théorique, n'épargne pas les établissements patrimoniaux, qui doivent souvent se contenter de projets de recherche innovants pour pouvoir profiter de davantage de personnel, alors que les tâches structurelles sont elles en souffrance. Un outil de gestion tel que Wikibase constitue donc un paradoxe intéressant : d'un côté, il représente l'opportunité d'améliorer, mais également de donner de la visibilité aux tâches de maintenance, d'un autre, de par son caractère novateur et expérimental, il constitue une autre de ces innovations, susceptible de ne pas être pérennisée et renfermant son propre lot d'opérations de maintenance, de sauvegarde et de mises à jour supplémentaires. Cependant, quelle que soit l'issue que connaîtra le prototype réalisé dans le cadre de cette thèse, nous pouvons d'ores et déjà relever qu'il aura eu le mérite de susciter des interrogations et réflexions s'étendant bien au-delà de son existence propre.

À la question « Comment favoriser une gestion soutenable des données d'autorité archivistiques dans le cadre du Web de données ? », nous proposons six éléments de réponses devant compléter les points déjà évoqués au cours des pages précédentes :

1. En réalisant un travail de sensibilisation, pour valoriser l'importance de l'indexation – une pratique jusque-là peu développée dans les ser-

62. À titre indicatif, la (nouvelle) Biographie nationale de Belgique (https://www.academieroyale.be/fr/la-biographie-nationale-personnalites/), qui constitue une source précieuse dans le cadre de la description de personnalités belges, est actuellement uniquement mise à disposition sous la forme de PDF de plusieurs centaines de pages correspondant à la version numérisée des exemplaires papier : nous sommes bien loin d'une mise à disposition de données structurées sur ces personnalités.

vices d'archives [63] – et de la production de métadonnées de qualité, comme briques nécessaires à l'élaboration de constructions plus ambitieuses [64]. Cela inclut également le fait de donner de la visibilité au travail de maintenance, d'ordinaire peu valorisé, afin que les institutions puissent disposer de moyen structurels adéquats, comme suggéré au cours des paragraphes précédents.

2. En établissant ce qui pourrait correspondre au degré de *good enough* (Greene et Meissner, 2005) dans le cadre de la description des personnes liées à des fonds d'archives : en effet, avec l'évolution des données d'autorité vers des entités pouvant être décrites de façon extrêmement riche et précise, la tentation est grande de vouloir stocker toute l'information imaginable, sous prétexte que cela peut être fait. Il est donc nécessaire de clarifier ce qui est considéré commme étant du ressort de l'institution dans le cadre d'une meilleure (re)contextualisation des collections et ce qui relève plutôt de la responsabilité du chercheur désireux d'approfondir ses connaissances.

3. En créant des outils et conditions permettant une étroite collaboration entre informaticiens, archivistes, historiens, mais également bibliothécaires, afin de créer du dialogue et améliorer l'intercompréhension. Bien que ce dialogue puisse s'avérer parfois délicat, il est fertile et nécessaire. De par leurs particularités les plaçant au carrefour entre ces différents profils, les données d'autorité pour les personnes physiques représentent ainsi un prétexte à de nouveaux types de pratiques collaboratives.

4. En adoptant des méthodes de travail *agiles* [65], à l'instar par exemple de ce que pratiquent the National Archives (UK) (Garmendia, 2019) ou les Archives nationales de France (Scalla, 2017), en commençant

63. Mais un vent de changement se fait sentir. Ainsi les Archives nationales de France ont par exemple décidé depuis 2018 d'utiliser un seul référentiel pour tous les agents, « quelles que soient ses relations avec les documents conservés aux AN. » (Clavaud et Charbonnier, 2020, p. 31).

64. Dans un contexte plus éloigné mais de façon similaire, l'une des promotrices du projet ADOCHS – dans le cadre duquel cette thèse a été réalisée – a publié un plaidoyer en faveur du financement des mathématiques comme « matière première de la révolution technologique », et pas seulement l'intelligence artificielle et la cybersécurité, voir : https://www.tijd.be/tijd/algemeen/investeer-in-de-grondstof-van-de-technologische-revolutie/10207038.

65. Nous reprenons ici la définition proposée par Scalla (2017) : « Utilisées d'abord dans le cadre du développement logiciel et désignant de nouvelles pratiques de gestion de projet, les méthodes agiles trouvent leur émergence dans les années 2000. S'imposant comme un modèle plus souple et plus flexible, elles prennent le contre-pied des gestions de projets dites prédictives en proposant le développement d'un produit par le biais d'itérations. » (Scalla, 2017, p. 17).

peut-être de façon modeste [66], mais dès maintenant, à investir le territoire que constitue le Web de données.

5. En favorisant une bonne documentation [67] – qui constitue un véritable enjeu quant à l'adoption d'un outil par les différents profils d'utilisateurs – mais également en veillant à sa mise à jour régulière [68]. L'enjeu est d'autant plus crucial que l'explication doit porter ici tant sur le contenu et sa modélisation [69], que sur un outil et ses fonctionnalités.

6. Enfin, en ayant l'honnêteté de se demander si le passage au RDF représente une réelle plus-value pour l'institution et ses utilisateurs ? Ce n'est en effet pas la solution à tout, comme le soulignent Lovins et Hillmann :

   > In the case of bibliographic vocabularies, there is sometimes a tendency toward wishful thinking, that the mere presence of Linked Open Data will solve the challenges of sustainability, scalability, and quality control ; that if one only had more RDF instance data and more controlled vocabularies, a stable ecosystem would emerge of its own accord. (Lovins et Hillmann, 2017)

   Par ailleurs, comme l'avance Gautier Poupeau, *data architect* à l'Institut National de l'Audiovisuel (INA) et expert du Web sémantique, l'usage de formats plus simples peut s'avérer préférable s'il permet une plus grande réutilisation des données (Poupeau, 2019).

   Il s'agira donc d'observer dans la pratique si les fonctionnalités offertes par l'interface collaborative Wikibase ainsi que la valeur ajoutée que peut représenter l'existence d'un point d'accès SPARQL justifient la complexité et la courbe d'apprentissage qu'elles impliquent.

66. Il s'agit également d'accepter certaines concessions : par exemple, au départ, notre plan visait à utiliser des éléments Wikibase plutôt que des chaînes de caractères pour les prénoms et noms, mais dans les faits, cela s'est avéré prématuré d'alourdir la charge avec cette complexité supplémentaire.

67. Comme le souligne Guillaud, une explication doit « aider l'utilisateur à comprendre les limites du système [...] Elle n'est pas un outil de communication qui doit faire disparaître ses biais, ses erreurs ou ses choix. Elle est un outil loyal qui doit montrer ses lacunes et ses failles » (Guillaud, 2019).

68. En effet, « l'explication est un processus continu. Fournir une explication n'est pas une fin en soi pas plus qu'[elle] n'est une chose que l'on fait une fois pour toutes : cela ne consiste pas à fournir un matériel didactique plus ou moins satisfaisant. Une explication sert à créer de la confiance à toutes les étapes des interactions, ce qui signifie que l'utilisateur doit être capable d'explorer activement les choix dont il dispose et notamment les erreurs possibles. » (Hoffman *et al.*, 2018, cité par Guillaud, 2019).

69. Or, étant donné toutes les exceptions caractérisant le réel, cela peut s'avérer fastidieux voire illusoire de chercher à englober tous les cas de figure susceptibles d'être rencontrés.

Perspectives

Finalement, bien que nous ayons cherché à envisager la gestion des données d'autorité avec la plus grande rigueur et toute l'acuité possible, nous sommes pourtant loin d'avoir épuisé notre objet d'étude. En effet, dans un contexte de métadonnées archivistiques en transition et d'évolution constante des modes de partage et réutilisation de données, cette thèse se présente avant tout comme une expérience réalisée – et documentée – à un instant T. Elle est donc destinée à être complétée et mise à jour à l'avenir par l'exploration d'axes de recherche connexes ou par l'approfondissement de pistes qui mériteraient davantage d'attention.

Tout d'abord, si notre attention s'est concentrée sur les personnes physiques et, indirectement sur les lieux liés à ces personnes, il serait intéressant de généraliser cette démarche à d'autres types d'entités, notamment les collectivités. En effet, elles représentent un défi stimulant dans le cadre de l'archivistique, dans la mesure où la structure et les fonctions de certains organismes sont amenées à évoluer à travers le temps, sans parler du fait que leurs dates d'existence sont parfois floues et que d'autres collectivités demeurent parfois les mêmes tout en étant appelées à changer de nom (Savoja et Vitali, 2008; Billinton, 2008). Les archivistes devant prendre en compte ces changements lors de la création de fichiers d'autorité, il serait donc intéressant de voir comment le modèle de triplets *augmentés* supportés par Wikibase peut prendre en charge et refléter ces réalités parfois complexes.

De façon similaire, il faut garder à l'esprit que l'initiative Records in Contexts entraîne dans son sillage des considérations plus globales : la question des référentiels et métadonnées archivistiques requiert de pouvoir être pensée plus largement qu'au seul niveau des personnes. Bien que ces dernières constituent à nos yeux une voie d'accès pertinente pour initier un changement progressif, elles gagneraient donc à être davantage réfléchies comme un type d'entités parmi d'autres, contenu au sein d'un écosystème global, et non d'une manière isolée.

Comme nous l'avons exposé au cours des chapitres précédents, l'utilisation du langage de requête SPARQL est requise pour pouvoir pleinement tirer parti des données structurées stockées dans une base de connaissance. Or, ce langage requiert une certaine littératie numérique [70]. Pour faire face

70. Comme le relate Lovink, l'informaticien Joseph Weizenbaum a mis en exergue *art of asking the right question* : « the problem of the Internet, according to Weizenbaum, is that it invites us to see it as a Delphic oracle. The Internet will provide the answer to all our questions and problems. But the Internet is not a vending machine in which you throw a coin and then get what you want. The key, here, is the acquisition of a proper education in order to formulate the right query. It's all about how one gets to pose the right question. For this one needs education and expertise » (Lovink, 2008).

aux lacunes des utilisateurs qui ne seraient pas familiers de ce type de langage de requête [71], deux approches peuvent être examinées. La première vise à pallier ces lacunes, tandis que la seconde se concentre sur l'élaboration de solutions alternatives. La première approche soulève dès lors la question du rôle que doivent jouer les archives, musées et bibliothèques dans la construction de ces nouveaux savoir-faire [72], tandis que la seconde approche se concentre sur le développement d'interfaces davantage adaptées à des êtres humains. En effet, une fois les données structurées, sémantisées et liées [73], des efforts supplémentaires sont requis pour que ces dernières puissent être facilement visualisées et prises en main par les utilisateurs. Ces efforts incluent tant la présentation de données dans un format plus convivial [74], que le développement d'outils permettant la traduction de questions formulées en langage naturel en requêtes SPARQL [75].

Comme souligné à plusieurs reprises, les possibilités de partage et de réutilisation de données entre bases de connaissance sont encore à leurs premiers balbutiements, au-delà des possibilités d'ores et déjà offertes par les requêtes fédérées. C'est donc un domaine de recherche qu'il vaudra la peine de continuer à explorer, au fur et à mesure que des solutions techniques se déploieront. Ces progrès devraient également être favorables aux démarches de réinjection de données dans Wikidata. La simplification de telles opérations devrait ainsi permettre de davantage concentrer les efforts sur le contenu, au-delà des questions de faisabilité. Ce faisant, les archives, musées et bibliothèques pourront par exemple investiguer la façon dont leurs données peuvent contribuer au manque de représentation et de diversité qui touche les projets Wikimédia (Wikimedia, 2018) et occasionne des problèmes tels que le *gender gap* (Klein *et al.*, 2016).

De telles contributions supposent toutefois un certain investissement. En effet, alors que les établissements patrimoniaux sont encouragés, tant par des acteurs ministériels (Filippetti, 2019), que par des acteurs de terrain (Molinié, 2020), à investir les projets Wikimédia, il faut prendre en consi-

71. Comme le relèvent Helmreich *et al.*, « most humanities researchers lack the skills to query these big data sets using sparql » (Helmreich *et al.*, 2019).

72. Ainsi, la chercheuse Ulrike Wuttke suggérait lors de l'édition 2018 de la conférence DH Benelux qu'il était du ressort des institutions, du moins des bibliothèques universitaires, de servir de centres d'expertise en *Data Science* à même d'offrir aux chercheurs des conseils sur la façon de produire et traiter les données (Cock *et al.*, 2018).

73. Une condition préalable, selon ce groupe de travail travaillant sur la visualisation des données dans le contexte des bibliothèques : « On ne peut rien faire sans des données normalisées, riches et liées » (Roux et Poveda, 2019).

74. C'est le cas par exemple de l'interface de découverte *(Passage Explorer)* développée par l'OCLC dans le cadre du projet Passage (Godby *et al.*, 2019, p. 25) ou encore de *Reasonator*, une interface de découverte basée sur les données Wikidata (Lemus-Roja et Pintscher, 2018).

75. Voir par exemple les travaux de Diefenbach *et al.* (2020) qui prennent en compte la question du multilinguisme.

dération le fait que cela requiert plus que de bonnes intentions [76]. Pour se familiariser avec l'univers Wikimédia et éviter d'en enfreindre des règles par ignorance, il peut donc être judicieux d'accueillir un wikimédien en résidence (Rey-Bellet, 2015) ou de nouer des collaborations et prendre conseil auprès d'un *chapitre* local Wikimédia (Yoakim, 2019). Cependant, il faut garder à l'esprit que de tels partenariats comportent également leurs zones d'ombres [77]. Les institutions sont donc encouragées « à mieux connaître l'environnement wikimédien, à avoir une stratégie, ainsi à se former et à avoir une implication forte au sein des projets, tout en favorisant une heureuse symbiose avec les bénévoles et leurs attentes » (Deshaye, 2019).

Par ailleurs, à l'heure où l'IFLA (International Federation of Library Associations and institutions), en partenariat avec quatre autres organisations internationales [78], alerte sur le fait qu'il est urgent d'agir, le patrimoine culturel mondial étant « menacé par les effets dévastateurs du changement climatique » (EIFL *et al.*, 2020, p. 1), il est nécessaire d'également adopter une perspective plus large. Il s'agira donc de poursuivre la recherche entreprise en réfléchissant à la forme que peut prendre une *gestion soutenable* non seulement à l'échelle de la donnée, mais également à l'échelle de l'environnement global dans lequel prennent place ces expérimentations. Comme le souligne Jackson :

> Broken world thinking asserts that breakdown, dissolution, and change, rather than innovation, development, or design as conventionally practiced and thought about are the key themes and problems facing new media and technology scholarship today. (Jackson, 2014, p. 222)

Tant le secteur des sciences de l'information que celui du patrimoine culturel sont donc appelés à mener une réflexion critique sur la façon dont cette

76. Molinié relève ainsi dans le contexte muséal que « s'approprier les outils et approcher la communauté wikimédienne peut susciter des appréhensions de la part des professionnels des musées, mêlant à la fois des questions de légitimité, de compétences techniques et de collaborations avec des personnes d'horizons souvent éloignés du domaine muséal » (Molinié, 2020). Des constats corroborés par Yoakim, qui va même jusqu'à utiliser les termes de *parcours initiatique* (Yoakim, 2019, p. 42).

77. Comme l'avait par exemple dénoncé Benoît Deshaye, wikimédien à l'origine de la plateforme Crotos : « si au cours des dix dernières années les avancées ont été notables, en particulier dans la reconnaissance institutionnelle et certains versements de contenus, il semblerait qu'aujourd'hui derrière les grand-messes, le bilan pourrait paraître bien pire que mitigé : des journées contributives sans lendemains, des partenariats dont on peine parfois à mesurer l'apport, des contenus majoritairement encore fermés [...], des contributeurs peu impliqués, des institutions absentes des projets, une célébration générale des projets Wikimédia qui ne se traduit pas en pratiques de contribution » (Deshaye, 2018).

78. L'International Council of Archives (ICA), l'International Council of Museums (ICOM), la Society of American Archivists (SAA) et l'Electronic Information for Libraries (EIFL).

mission de mise à disposition des données peut continuer à prendre place, en tenant compte des enjeux environnementaux actuels.

Enfin, pour conclure, nous nous référerons à cette expression de Philippe Le Pape, qui prônait, dans le cadre de la transition bibliographique, le refus d'un *grand soir catalographique*, au profit de *nombreux petits matins* (Le Pape, 2015, cité par Cavalié, 2019). Comme nous avons cherché à le montrer dans cette thèse mettant l'accent sur la réalité du terrain, nous adhérons pleinement à cette vision favorable aux évolutions progressives et aux petits pas discrets, réalisés avec constance et persévérance dans la pénombre du petit matin – pour peu toutefois que les moyens suivent.

Bibliographie

Bibliographie

AALTO UNIVERSITY (2017). WarSampo won the Open Data Prize in the 2017 LODLAM Challenge. [En ligne], https://www.aalto.fi/en/news/warsampo-won-the-open-data-prize-in-the-2017-lodlam-challenge, consulté le 30.09.2020.

ABES (2019). FNE - Preuve de Concept en cours. [En ligne], https://fil.abes.fr/2019/09/04/fne-preuve-de-concept-en-cours, consulté le 23.04.2020.

ABES ET BNF (2019a). Inscriptions ouvertes pour la 4e journée professionelle « Métadonnées en bibliothèques » du 15 novembre. [En ligne], https://www.transition-bibliographique.fr/2019-09-26-inscriptions-ouvertes-4e-journee-metadonnees-bibliotheques-15-novembre-2019/, consulté le 29.08.2020.

ABES ET BNF (2019b). Les rôles et les restrictions dans Wikibase. [En ligne], https://github.com/abes-esr/poc-fne/issues/211, consulté le 23.05.2020.

ABES ET BNF (2020). Le Fichier national d'entités (FNE) : décryptage. [En ligne], https://www.transition-bibliographique.fr/fne/fichier-national-entites/, consulté le 23.04.2020.

ABIÁN, D., GUERRA, F., MARTÍNEZ-ROMANOS, J. et TRILLO-LADO, R. (2017). Wikidata and DBpedia : A Comparative Study. *In Semantic Keyword-based Search on Structured Data Sources*, pages 142–154. Springer.

ADAMCZEWSKI, G. (1988). La recherche-action. *Recherche & formation*, 3(1): 109–114.

AERTS, K., LUYTEN, D., WILLEMS, B. et DROSSENS, P. (2017). *Papy était-il un nazi ? Sur les traces d'un passé de guerre*. Éditions Racine, Bruxelles.

AEYERS, P., JANSSEN, O. et ANON. (2019). Wikidata - Wikibase interest meeting at the National Library of Sweden. Notes. [En ligne], https:

//web.archive.org/web/20200428093246/https://docs.google.com/document/d/1ZYq5-H54K671zdqAoFUgROPzZsT4cfyP11nbdCS_KUw/edit, consulté le 28.04.2020.

ALGERGAWY, A., FARIA, D., FERRARA, A., FUNDULAKI, I., HARROW, I., HERTLING, S., JIMÉNEZ-RUIZ, E., KARAM, N., KHIAT, A., LAMBRIX, P. *et al.* (2019). Results of the ontology alignment evaluation initiative 2019. *In CEUR Workshop Proceedings*, volume 2536, pages 46–85.

ALLISON-CASSIN, S., ARMSTRONG, A., AYERS, P., CRAMER, T., CUSTER, M., LEMUS-ROJAS, M., MCCALLUM, S., PROFFITT, M., PUENTE, M., RUTTENBERG, J. *et al.* (2019). ARL White Paper on Wikidata : Opportunities and Recommendations.

ALLISON-CASSIN, S. et SCOTT, D. (2018). Wikidata : a platform for your library's linked open data. *Code4Lib Journal*, (40).

ALLISON-CASSIN, S. et SEEMAN, D. (2019). Leveraging Wikibase for Linked Data Vocabulary Management : Indigenous Communities in Canada. *In 2019 LD4 Conference on Linked Data in Libraries*. [En ligne], https://docs.google.com/presentation/d/1gD0j304DAdqE4-QM3SG2gvWhmWJwKigmrA8Heh-MbeY/edit, consulté le 31.05.2019.

ALVAREZ, A. (2019). Multilingual Wikidata Infoboxes. An opportunity to a painless migration for any Wikipedia. Intervention dans le cadre de la WikidataCon 2019, [En ligne], https://commons.wikimedia.org/wiki/File:WikidataCon-2019._Multilanguage_Infoboxes._Amador_Alvarez.pdf, consulté le 15.05.2020.

ANDERSON, JILLIAN (2018). Record comparison with recordlinkage. [En ligne], https://uwaterloo.ca/networks-lab/blog/post/record-pair-classification-recordlinkage, consulté le 02.09.2020.

ANGJELI, A. et BOBER, B. (2019). Assessing Wikibase as the core for the French National Entities File (FNE). Intervention dans le cadre de la WikidataCon 2019, [En ligne], https://commons.wikimedia.org/w/index.php?title=File:Wikibase_for_FNE.pdf, consulté le 15.05.2020.

ANGJELI, A., CLAVAUD, F. et ROUSSEL, S. (2017). Représenter en rdf, interconnecter et visualiser en graphe des jeux de métadonnées archivistiques de provenances multiples : un projet de prototype. *Gazette des archives*, 245(1):157–171.

ANTONIN, D. (2018). Extend Wikidata extension to support arbitrary Wikibase instances. [En ligne], https://github.com/OpenRefine/OpenRefine/issues/1640#issuecomment-395695006, consulté le 24.05.2020.

ANTRACOLI, A. A. et RAWDON, K. (2019). What's in a Name ? Archives for Black Lives in Philadelphia and the Impact of Names and Name Authorities in Archival Description. *In* SANDBERG, J., éditeur : *Ethical Questions in Name Authority Control*, pages 307–336. Library Juice Press.

ARCHIVES DE L'ÉTAT (2020). Plus de mille inventaires « rétroconvertis » ont été téléchargés pendant le confinement ! [En ligne], http://arch.arch.be/index.php?l=fr&m=actualites&r=toutes-les-actualites&a=2020-05-27-plus-de-mille-inventaires-retroconvertis-ont-ete-telecharges-pendant-le-confinement, consulté le 29.09.2020.

ARCHIVES NATIONALES DE FRANCE (2018). Résultats du hackathon des Archives nationales. [En ligne], http://www.archives-nationales.culture.gouv.fr/resultats-du-hackathon-des-archives-nationales, consulté le 21.04.2020.

ARCHIVES NATIONALES DE FRANCE (2019). Journée d'étude du 28 janvier 2020 Pierrefite-sur-Seine. Les métadonnées archivistiques en transition. [En ligne], https://francearchives.fr/file/d7d67d46e1e6316b4b66bb84ded77d50df9b4cce/programme_28janvier2020-def.pdf, consulté le 29.08.2020.

ARNOLD, H. (2016). Critical Work : Archivists as Maintainers. [En ligne], https://hillelarnold.com/blog/2016/08/critical-work, consulté le 12.02.2020.

ARNOLD, K. (2019). Eac-cpf revision half-way into stage 2. [En ligne], https://archivesportaleurope.blog/2019/12/19/eac-cpf-revision-half-way-into-stage-2/, consulté le 01.10.2020.

@B2C (2019). #Wikibase est une solution évidente pour tout futur logiciel #LOD de description/gestion d'#archives. Un de ces principaux avantages (parmi tant d'autres) est de pouvoir clairement sourcer les déclarations. Les autres : interopérabilité, outils liés, etc. [Tweet], https://web.archive.org/web/20191120231049/https:/twitter.com/b3d2c/status/1174265202105356290.

BÁNKI, Z., MÉSZÁROS, T., NÉMETH, M. et SIMON, A. (2016). Checking the identity of entities by machine algorithms : The next step to the Hungarian National Namespace. *Code4Lib Journal*, (33).

BARATS, C. (2013). *Manuel d'analyse du web en sciences humaines et sociales*. Armand Colin.

BARATS, C., LEBLANC, J.-M. et FIALA, P. (2013). Approches textométriques du web : corpus et outils. *Manuel d'analyse du web en Sciences Humaines et Sociales*, pages 100–124.

BARTHOLMEI, S., FRANKS, R., HEILMAN, J., JOSEPH, M., MCDONALD, V., RAUNIK, A., RIDGE, M. et ROBERTSON, M. (2016). Opportunities for Academic and Research Libraries and Wikipedia. [En ligne], https://www.ifla.org/files/assets/hq/topics/info-society/iflawikipediaopportunitiesforacademicandresearchlibraries.pdf, consulté le 03.06.2020.

BARTHÉLÉMY, D. (5 Décembre 2018). Kale, Kiwi... De plus en plus de bébés ont des prénoms d'aliments sains. *Slate FR*. [En ligne], https://www.slate.fr/story/170850/kale-kiwi-prenoms-bebes-aliments-sains?fbclid=IwAR0W7OMLPCvOKfbmwoufUYiE20XQrm08SKmO01zSft5qnGCfdXa9UdvPNMs, consulté le 10.01.2019.

BASKAUF, S. (2019). Putting Data into Wikidata using Software. [En ligne], http://baskauf.blogspot.com/2019/06/putting-data-into-wikidata-using.html, consulté le 17.07.2020.

BASS, S. (2009). How Many Different Ways Can You Spell 'Gaddafi'. *ABC News*. [En ligne], https://web.archive.org/web/20120206125143/http:/abcnews.go.com/blogs/headlines/2009/09/how-many-different-ways-can-you-spell-gaddafi/, consulté le 26.08.2020.

BATAILLE, M. (1983). Méthodologie de la complexité. *Revue Pour*, (90):32–36.

BEARMAN, D. (1989). Authority control issues and prospects. *The American Archivist*, 52(3):286–299.

BEARMAN, D. A. et LYTLE, R. H. (1985). The power of the principle of provenance. *Archivaria*, 21:14–27.

BECKER, J. (2017). Indexing candidate links with recordlinkage. [En ligne], https://uwaterloo.ca/networks-lab/blog/post/indexing-candidate-links-recordlinkage, consulté le 02.09.2020.

BECOMPTA (2020a). Personne morale. [En ligne], https://www.becompta.be/dictionnaire/personne-morale, consulté le 30.08.2020.

BECOMPTA (2020b). Personne physique. [En ligne], https://www.becompta.be/dictionnaire/personne-physique, consulté le 30.08.2020.

BÉRANGER, L. (2017). Indexation et catalogue informatique du centre de documentation cegesoma : analyse de l'existant et pistes d'amélioration. Mémoire de D.E.A., Université catholique de Louvain, Louvain-la-Neuve.

BERNERS-LEE, T. (1998). Cool URIs don't change. [En ligne], https://www.w3.org/Provider/Style/URI, consulté le 12.12.2020.

BERNERS-LEE, T., FIELDING, R. et MASINTER, L. (2005). Uniform Resource Identifier (URI) : Generic Syntax. RFC 3986, [En ligne], https://tools.ietf.org/html/rfc3986, consulté le 30.09.2020.

BEST VALUE (2003). Best Value and Local Authority Archives. [En ligne], https://web.archive.org/web/20040606100505/http://www.bestvaluefororarchives.org.uk/competition.htm, consulté le 31.08.2020.

BEVILACQUA, S. (2016). Corpus web 2.0 : quelques enjeux méthodologiques et épistémologiques. *Synergies Argentine*, (4):81–93.

@BIBLIOQC (2020). Aux @bibliomontreal, on utilise bibliotechnicien-ne-s et aux @bibUdeM, @BiblioUQAM, @_BAnQ on retrouve technicien-ne-s en documentation. Les deux s'utilisent au Québec et au Canada mais le 2e est plus commun (me semble). [Tweet], https://web.archive.org/web/20200224133225/https://twitter.com/BiblioQC/status/1231930413028958209.

BILLINTON, S. (2008). The role of archival authority records in the finding aid system of the archives of ontario. *Journal of Archival Organization*, 5(1-2):75–93.

BIRKHOLZ, J. (2020). Decomplexifying the network pipeline : a tool for rdf/wikidata to network analysis. *DH Benelux Journal*, 2.

BJÖRK, B.-C. et SOLOMON, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of informetrics*, 7(4):914–923.

BLANCHET, P. (2000). *La linguistique de terrain, méthode et théorie. Une approche ethno-sociolinguistique*. Presses Universitaires de Rennes.

BNF (2018). NOEMI : vers un nouvel outil de production des métadonnées de la BnF. [En ligne], https://www.bnf.fr/fr/noemi-vers-un-nouvel-outil-de-production-des-metadonnees-de-la-bnf, consulté le 23.04.2020.

BOONSTRA, O., BREURE, L. et DOORN, P. (2004). Past, present and future of historical information science. *Historical Social Research/Historische Sozialforschung*, pages 4–132.

BOULLIER, D. et LOHARD, A. (2012). *Opinion mining et Sentiment analysis : Méthodes et outils*. OpenEdition Press.

BOURDELOIE, H. (2014). Ce que le numérique fait aux sciences humaines et sociales. épistémologie, méthodes et outils en questions. *tic&société [Online]*, 7(2). [En ligne], http://journals.openedition.org/ticetsociete/1500, consulté le 10.06.2020.

BOURDON, F. (1997). Qu'est-ce qu'un format d'autorité. *Bulletin d'information de l'Association des bibliothécaires français*, (175):46–52.

BOYDENS, I. (2001). Déploiement coopératif d'un dictionnaire électronique de données administratives. *Document numérique*, 5(3):27–43.

BOYDENS, I. et van HOOLAND, S. (2011). Hermeneutics applied to the quality of empirical databases. *Journal of documentation*, 67(2):279–289.

BRAZZO, L. et MAZZINI, S. (2015). Open memory project. [En ligne], https://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf, consulté le 30.09.2020.

BROWN, S. et SIMPSON, J. (2013). The curious identity of michael field and its implications for humanities research with the semantic web. *In 2013 IEEE International Conference on Big Data*, pages 77–85. IEEE.

BRUCE, T. R. et HILLMANN, D. I. (2004). The Continuum of Metadata Quality : Defining, Expressing, Exploiting. *In Metadata in Practice*. ALA editions.

BYRNE, W. et WYATT, L. (2019). Wikidata Wikibase for National Libraries : the inaugural meeting. [En ligne], https://pro.europeana.eu/post/wikidata-wikibase-for-national-libraries-the-inaugural-meeting, consulté le 23.04.2020.

BYRUM JR, J. D. (2004). NACO : A Cooperative Model for Building and Maintaining a Shared Name Authority Database. *Cataloging & Classification Quarterly*, 38(3-4):237–249.

CASALINI, M., CHEW, C. N., CLUFF, C., DUROCHER, M., FOLSOM, S., FRANK, P., GATENBY, J., GODBY, J., KOVARI, J., LORIMER, N. *et al.* (2018). National Strategy for Shareable Local Name Authorities National Forum : White Paper. White paper. [En ligne], https://ecommons.cornell.edu/handle/1813/56582, consulté le 10.09.2020.

CAVALIÉ, E. E. (2019). *L'indexation matière en transition : De la réforme de Rameau à l'indexation automatique. Sous la direction de Étienne Cavalié.* Ed. du Cercle de la librairie.

CEGESOMA (2007). Le Mission Statement du CegeSoma. *Bulletin du CegeSoma*, (40):5–6. [En ligne], https://www.cegesoma.be/docs/media/Bulletins/Bulletin40.pdf, consulté le 06.08.2020.

CEGESOMA (2015). Cegesoma - gestion des métadonnées. rapport d'audit 2015. Rapport, CegeSoma.

CEGESOMA (2018a). Plan opérationnel du CegeSoma 2018-2021. Version 1. Rapport, CegeSoma.

CEGESOMA (2018b). Rapport annuel 2017. Rapport, CegeSoma.

CEGESOMA (2019a). Mission statement. [En ligne], http://www.cegesoma.be/docs/media/Divers/missionstatement_fr.pdf, consulté le 24.05.2019.

CEGESOMA (2019b). Rapport annuel 2018. Rapport technique, CegeSoma.

CEGESOMA (2020a). Le CegeSoma. Histoire du Centre. [En ligne], https://www.cegesoma.be/fr/le-cegesoma, consulté le 06.08.2020.

CEGESOMA (2020b). Mission Organisation du CegeSoma. [En ligne], https://www.cegesoma.be/fr/mission-organisation-du-cegesoma, consulté le 06.08.2020.

CEGESOMA (2020c). Nos bénévoles. [En ligne], https://www.cegesoma.be/fr/nos-bénévoles, consulté le 06.08.2020.

CEGESOMA (2020d). Nouveaux projets de recherche. [En ligne], https://www.cegesoma.be/fr/nouveaux-projets-de-recherche, consulté le 16.08.2020.

CENTRE NATIONAL DE RESSOURCES TEXTUELLES ET LEXICALES (2020). Soutenable. [En ligne], https://www.cnrtl.fr/definition/soutenable, consulté le 02.09.2020.

CHARDONNENS, A. (2017). Les données web analytics. Rapport, MADDLAIN Project. [En ligne], https://www.cegesoma.be/docs/images/stories/cegesoma/Recherche/Maddlain_WebAnalytics.pdf, consulté le 19.07.2020.

CHARDONNENS, A. et HENGCHEN, S. (2017). Text mining for user query analysis : A 5-step method for cultural heritage institutions. *In Proceedings of the 15th International Symposium on Information Science (ISI 2017) ; Berlin, Germany, 13th—15th March 2017 : Everything Changes, Everything Stays the Same ? Understanding Information Spaces*, pages 177–189. M. Gäde/V. Trkulja/V. Petras (Eds.).

CHARDONNENS, A., HUNGENAERT, J. et VANBRABANT, M. (2017). Combiner analyses quantitatives et qualitatives afin de mieux comprendre les pratiques et besoins des utilisateurs des archives et bibliothèques. *In Actes de la journée d'étude Inside the User's Mind, organisée dans le cadre du projet MADDLAIN, le 22 février 2017*, pages 11–23.

CHARDONNENS, A., RIZZA, E., COECKELBERGS, M. et VAN HOOLAND, S. (2018). Mining User Queries with Information Extraction Methods and Linked Data. *Journal of Documentation*, 74(5):936–950.

CHARTIER, D. (2013). Enjeux méthodologiques de l'étude contemporaine de l'actualité dans la presse : le cas de l'image de l'Islande pendant la crise économique. *Médias 19*, (42). [En ligne], http://www.medias19.org/index.php?id=15543, consulté le 04.06.2020.

CHEIN, I., COOK, S. W. et HARDING, J. (1948). The field of action research. *American Psychologist*, 3(2):43–50.

CHRISTEN, D. M. (2012). *Concepts and Techniques for Record Linkage*. Springer.

CLAIR, K. (2016). Technical debt as an indicator of library metadata quality. *D-Lib Magazine*, 22(11):3.

CLAVAUD, F. (2019a). ICA-Records in Contexts conceptual model and ontology. Intervention dans le cadre de la journée d'étude Linking the Past (22 novembre 2019), [En ligne], http://adochs.be/wp-content/uploads/2020/01/LinkingThePast_Brussels_20191122_RecordsInContexts.pdf, consulté le 12.09.2020.

CLAVAUD, F. (2019b). Transformer les métadonnées des Archives nationales en graphe de données : enjeux et premières réalisations. *La Gazette des archives*, 254(2):59–88.

CLAVAUD, F. (2020a). Le nouveau standard international de description des archives *Records in Contexts*. Intervention dans le cadre de la Journée d'étude du 28 janvier 2020 aux Archives nationales, [En ligne], https://f.hypotheses.org/wp-content/blogs.dir/2167/files/2020/02/20200128_2_RecordsInContexts.pdf, consulté le 12.12.2020.

CLAVAUD, F. (2020b). Ric-o converter : un exemple de mise en application du standard ica records in contexts (ica ric). [En ligne], https://blog-ica.org/fr/2020/06/13/ric-o-converter-un-exemple-de-mise-en-application-du-standard-ica-records-in-contexts-ica-ric/, consulté le 06.09.2020.

CLAVAUD, F. et CHARBONNIER, P. (2020). Records in contexts aux archives nationales : enjeux et premières réalisations. Intervention dans le cadre de la journée d'étude Les métadonnées archivistiques en transition (28 janvier 2020), [En ligne], https://f.hypotheses.org/wp-content/blogs.dir/2167/files/2020/02/20200128_3_RiCauxAN_EnjeuxPremieresRealisations.pdf, consulté le 06.09.2020.

CLAVAUD, F. et CHÂTEAU-DUTIER, E. (2017). Une Preuve de Concept pour la Sémantisation et la Visualisation Orientée Utilisateur de Donnéees Archivistiques. *In 12th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2017, Montréal, Canada, August 8-11, 2017, Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO). [En ligne], https://dh2017.adho.org/abstracts/396/396.pdf, consulté le 06.04.2020.

CLAVAUD, FLORENCE (2018). Sémantisation et visualisation de métadonnées archivistiques : mise en ligne du prototype français PIAAF. [En ligne], https://www.ica.org/fr/semantisation-et-visualisation-de-metadonnees-archivistiques-mise-en-ligne-du-prototype-francais, consulté le 06.10.2020.

CLEVE, A. (2016). Analyzing the Evolution of Data-Intensive Software Systems in Support to Software Maintenance. [En ligne], http://mastic.ulb.ac.be/wp-content/uploads/2016/02/cleve.pdf, consulté le 31.08.2020.

COCK, M., ten DOLLE, E. et CLAEYSSENS, S. (2018). Integrating Libraries and Digital Humanities. *In DH Benelux 2018*.

COLIGNON, A. (2018). Note de vision 2018-2019. Note de vision, Bibliothèque du CegeSoma.

COLIGNON, A. (2019). La bibliothèque du CegeSoma, une approche critique et personnelle. *Revue Belge d'Histoire Contemporaine*, XLIX(2-3):149–159.

COMITÉ DES NORMES ET BONNES PRATIQUES (2019). Rapport d'étape pour la révision et l'harmonisation des normes de description de l'ICA. [En ligne], https://www.ica.org/sites/default/files/Rapport_pour%20la%20r%C3%A9vision_harmonisation_normes_de_description.pdf, consulté le 11.08.2020.

@CONTEMPLATINGIM (2018). Interesting : https://arbido.ch/fr/edition-article/2018/automatisierung-versprechen-oder-drohung/archives-et-wikidata Am slowly concluding that software such as Wikibase not just useful for archives but also has potential applications in recordsmanagement and

corporate infomanagement more generally ... nextgenrm ontology taxonomy. [Tweet], https://web.archive.org/web/20191121172904/https://twitter.com/contemplatingIM/status/1039841111475335168.

COSENTINO, V., IZQUIERDO, J. L. C. et CABOT, J. (2015). Assessing the bus factor of git repositories. *In 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pages 499–503. IEEE.

CSI MINES PARISTECH (2016). Investigating maintenance and repair. [En ligne], http://www.csi.mines-paristech.fr/en/featured-articles/investigating-the-maintenance-and-repair-activity/?lang=en, consulté le 02.09.2020.

CUNNINGHAM, W. (1992). The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*, 4(2):29–30.

DAGIRAL, É. et PEERBAYE, A. (2012). Les mains dans les bases de données. *Revue d'anthropologie des connaissances*, 6(1):191–216.

DASU, T. et JOHNSON, T. (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons.

DATI.CDEC (2015). Shoah Vocabulary Specification. [En ligne], http://dati.cdec.it/lod/shoah/reference-document.html, consulté le 30.09.2020.

DAVIES, W. (Thursday 19 January 2017). How statistics lost their power - and why we should fear what comes next. *The Guardian*. [En ligne], https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy, consulté le 14.04.2017.

@DAVISKELLYK (2019). how do you justify the need for funding research and unsexy metadata work ? I get it that it doesn't appeal to the higher ups. but it. is. NECESSARY. work. you want sexy platform ? you need lots of unsexy data cleaning. [Tweet], https://twitter.com/daviskellyk/status/1159160602440323073.

DE WILDE, M. (2020). I want to talk to a human. Impact de la qualité des bases de connaissances sur les agents conversationnels. [En ligne (PDF)], http://mastic.ulb.ac.be/wp-content/uploads/2020/01/fnrs2020mdw.pdf, consulté le 11.05.2020.

DELPEUCH, A. (2020). Wikibase reconciliation interface for OpenRefine. [En ligne], https://github.com/wetneb/openrefine-wikibase/blob/master/README.md, consulté le 13.09.2020.

DEMAZIÈRE, D., HORN, F. et ZUNE, M. (2006). Dynamique de développement des communautés du logiciel libre. *Terminal 97*, 98:71–84.

DENIS, J. et GOËTA, S. (2013). La fabrique des données brutes. le travail en coulisses de l'open data. *In Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques*.

DENIS, J. et PONTILLE, D. (2014). Maintenance work and the performativity of urban inscriptions : the case of Paris subway signs. *Environment and Planning D : Society and Space*, 32(3):404–416.

DEPOORTERE, R. et DE SCHAMPHELAERE, F. (2019). Visienota. Hergebruik overheidsinformatie in het Algemeen Rijksarchief. Note de vision, Archives de l'État.

DEPOORTERE, R., GILLET, F. et ROEGES, M. (2018). Rapport des ateliers de la journée du 05 février 2018. Le numérique aux Archives de l'Etat pour répondre aux besoins des Universités. Rapport, Les Archives de l'État.

DESHAYE, B. (2018). WikiConvention francophone 2017 - Programme GLAM-Wiki : Qu'est-ce qui ne va pas ? - Et surtout, comment ça pourrait aller mieux ? [En ligne], https://meta.wikimedia.org/wiki/WikiConvention_francophone/2017/Programme/GLAM-Wiki_-_Qu'est-ce_qui_ne_va_pas_?_\T1\textendash_Et_surtout,_comment_ça_pourrait_aller_mieux, consulté le 20.05.2020.

DESHAYE, B. (2019). Wikimédia France - Assemblée générale 2018. Benoît Deshaye. [En ligne], https://meta.wikimedia.org/wiki/Wikimédia_France/Assemblée_générale/2018/Benoît_Deshayes, consulté le 20.05.2020.

DESMET, G. (2019). Verwerving en waadering in het cegesoma. visienota. Note de vision, CegeSoma.

DIEFENBACH, D., BOTH, A., SINGH, K. et MARET, P. (2020). Towards a question answering system over the semantic web. *Semantic Web*, (Preprint):1–19.

DIGITAL AND POPULATION DATA SERVICES AGENCY (27 août 2020). Changing surname. [En ligne], https://dvv.fi/en/changing-surname, consulté le 27.08.2020.

DOUGLAS, J., BAK, G., MCLELLAN, E., van HOOLAND, S. et FROGNER, R. (2018). Decolonizing archival description : Can linked data help ? *Proceedings of the Association for Information Science and Technology*, 55(1):669–672.

DOWNEY, G. J. (2014). Making media work : Time, space, identity, and labor in the analysis of information and communication infrastructures. *Media technologies : Essays on communication, materiality, and society*, pages 141–66. [En ligne], https://gdowney.files.wordpress.com/2013/11/downey-g-2014-in-gillespie-t-et-al-eds-2014-making-media-work.pdf, consulté le 12.02.2020.

DRABINSKI, E. (2013). Queering the Catalog : Queer Theory and the Politics of Correction. *The Library Quarterly*, 83(2):94–111.

DUCHARME, B. (2018). Running and querying my own Wikibase instance. [En ligne], http://www.snee.com/bobdc.blog/2018/06/running-and-querying-my-own-wi.html, consulté le 21.04.2020.

DULAURANS, M. (2015). CIFRE : parcours de compétences d'une thèse annoncée. *Revue française des sciences de l'information et de la communication*, (6).

DUSETZINA, S. B., TYREE, S., MEYER, A.-M., MEYER, A., GREEN, L. et CARPENTER, W. R. (2014). An overview of record linkage methods. *In Linking Data for Health Services Research : A Framework and Instructional Guide [Internet], note=[En ligne], https://www.ncbi.nlm.nih.gov/books/NBK253312/, consulté le 11.09.2020*. Agency for Healthcare Research and Quality (US).

EAD (EAD-BIBLIOTHEQUE.FR) (2020a). EAD en bibliothèque, guide des bonnes pratiques. Biographie ou histoire <bioghist>. [En ligne], https://www.ead-bibliotheque.fr/guide/contexte/bioghist/, consulté le 02.09.2020.

EAD (EAD-BIBLIOTHEQUE.FR) (2020b). EAD en bibliothèque, guide des bonnes pratiques. Origine - origination. [En ligne], https://www.ead-bibliotheque.fr/guide/donnees-du-did/origination/, consulté le 02.09.2020.

EAD (EAD-BIBLIOTHEQUE.FR) (2020c). EAD en bibliothèque, guide des bonnes pratiques. Vedettes et accès contrôlés <controleaccess>. [En ligne], https://www.ead-bibliotheque.fr/guide/indexation/, consulté le 02.09.2020.

EHRI (2020). EHRI Partners. [En ligne], https://www.ehri-project.eu/consortium, consulté le 28.09.2020.

EIDSON, J. G. et ZAMON, C. J. (2019). EAD Twenty Years Later : A Retrospective of Adoption in the Early Twenty-first Century and the Future of EAD. *The American Archivist*, 82(2):303–330.

EIFL, ICA, ICOM, IFLA et SAA (2020). Journée mondiale de la propriété intellectuelle - 26 avril 2020. Climat, patrimoine et propriété intellectuelle. [En ligne], https://www.ifla.org/files/assets/clm/news/wipo_letter_in_french_-_world_ip_day_2020.pdf, consulté le 12.09.2020.

ENSMENGER, N. (2016). When Good Software Goes Bad : The Unexpected Durability of Digital Technologies. *In Maintainers Conference, April 9, 2016*.

EQUIPE PROJET PIAAF (2018a). Enjeux, objectifs et historique du projet. PIAAF (Pilote d'interopérabilité pour les Autorités Archivistiques françaises) : démonstrateur. Archives nationales - Bibliothèque nationale de France - Service interministériel des Archives de France - société Logilab, [En ligne], https://piaaf.demo.logilab.fr/editorial/historique, consulté le 06.04.2020.

EQUIPE PROJET PIAAF (2018b). Réalisation. PIAAF (Pilote d'interopérabilité pour les Autorités Archivistiques françaises) : démonstrateur. Archives nationales - Bibliothèque nationale de France - Service interministériel des Archives de France - société Logilab, [En ligne], https://piaaf.demo.logilab.fr/editorial/contexte-technique, consulté le 06.04.2020.

ERXLEBEN, F., GÜNTHER, M., KRÖTZSCH, M., MENDEZ, J. et VRANDEČIĆ, D. (2014). Introducing Wikidata to the Linked Data Web. *In International Semantic Web Conference*, pages 50–65. Springer.

ESTERMANN, B. (2018). How Wikidata Is Solving Its Chicken-or-Egg-Problem in the Field of Cultural Heritage. [En ligne], https://www.societybyte.swiss/2018/11/07/how-wikidata-is-solving-its-chicken-or-egg-problem-in-the-field-of-cultural-heritage, consulté le 15.05.2020.

ESTERMANN, B., GSCHWEND, A., HALLER, S. et PARRALES MACHUCA, E. M. (2020). Basisregister und kontrollierte Vokabulare als Wegbereiter für Linked Open Data in der Schweiz. Berner Fachhochschule, Institut Public Sector Transformation, [En ligne], https://arbor.bfh.ch/10249, consulté le 30.09.2020.

EUROPEANA (2017). Get your vocabularies in Wikidata... [En ligne], https://pro.europeana.eu/page/get-your-vocabularies-in-wikidata, consulté le 03.06.2020.

FACTGRID (2019). FactGrid : Troubleshooting. [En ligne], https://database.factgrid.de/w/index.php?title=FactGrid:Troubleshooting&oldid=1662570, consulté le 27.04.2020.

FACTGRID (2020a). FactGrid : Troubleshooting. [En ligne], https://database.factgrid.de/w/index.php?title=FactGrid:Troubleshooting&oldid=2460210, consulté le 22.05.2020.

FACTGRID (2020b). FactGrid :Terms of Service. [En ligne], https://database.factgrid.de/wiki/FactGrid:Terms_of_Service, consulté le 27.04.2020.

FÄRBER, M., BARTSCHERER, F., MENNE, C. et RETTINGER, A. (2018). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129.

FAUCONNIER, S. (2018). Many faces of Wikibase : Rhizome's archive of born-digital art and digital preservation. [En ligne], https://wikimediafoundation.org/news/2018/09/06/rhizome-wikibase/, consulté le 21.04.2020.

FERNANDEZ, J. (2009). A SWOT analysis for social media in libraries. *Online*, 33(5):35–37.

FILIPPETTI, A. (2019). Lancement DBpedia en français et inaguration de Semanticpédia. [En ligne], https://www.culture.gouv.fr/Presse/Archives-Presse/Archives-Discours-2012-2018/Annee-2012/Lancement-DBpedia-en-francais-et-inauguration-de-Semanticpedia, consulté le 20.05.2020.

FISCHER, B. (2018). Authority Control meets Wikibase. [En ligne], https://wiki.dnb.de/display/GND/Authority+Control+meets+Wikibase, consulté le 21.04.2020.

FISCHER, B. et MANECKE, M. (2019). Über die allmählche Verfertigung der Gedanken beim Reden. [En ligne], https://wiki.dnb.de/pages/viewpage.action?pageId=148603894, consulté le 29.09.2020.

FISCHER, B. et OHLIG, J. (2019). New testing ground for Wikibase : A federal agency goes on an expedition in the Wiki universe. [En ligne], https://web.archive.org/web/20190509104451/https://blog.wikimedia.de/2019/05/09/new-testing-ground-for-wikibase-a-federal-agency-goes-on-an-expedition-in-the-wiki-universe/, consulté le 21.04.2020.

FISCHER, B. et OHLIG, J. (2020). Report "GND meets Wikibase" 2. Could you wikify an authority file ? [En ligne], https://wiki.dnb.de/pages/viewpage.action?pageId=167019461, consulté le 21.04.2020.

FLYVBJERG, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2). 219–245.

FOLI, O. et DULAURANS, M. (2013). Tenir le cap épistémologique en thèse Cifre. Ajustements nécessaires et connaissances produites en contexte. *Études de communication*, (1):59–76.

FRANAR (2010). Fonctionnalités requises des données d'autorité. Rapport final du Groupe de travail IFLA sur les Fonctionnalités requises et la numérotation des notices d'autorité (FRANAR). [En ligne], https://www.ifla.org/files/assets/cataloguing/frad/frad_2009-fr.pdf, consulté le 19.07.2020.

FRANCART, T. et CHARBONNIER, P. (2020). RIC-O converter, un logiciel libre de conversion de métadonnées archivistiques (en EAD et EAC-CPF) en jeux de données conformes à RiC-O. Intervention dans le cadre de la journée d'étude Les métadonnées archivistiques en transition (28 janvier 2020), [En ligne], https://f.hypotheses.org/wp-content/blogs.dir/2167/files/2020/02/20200128_4_RiCOConverter.pdf, consulté le 06.09.2020.

FREIRE, N. et ISAAC, A. (2019). Technical Usability of Wikidata's Linked Data. *In International Conference on Business Information Systems*, pages 556–567. Springer.

GAGNON, Y.-C. (2005). *L'étude de cas comme méthode de recherche : guide de réalisation*. Presses de l'Université du Québec.

GARMENDIA, J. (2019). Digital description and metadata at the national archives. digital strategy. *ABB : Archives et Bibliothèques de Belgique*, 106: 103–110.

GDPR.EU (2020). Recital 27. Not applicable to data of deceased persons. [En ligne], https://gdpr.eu/recital-27-not-applicable-to-data-of-deceased-persons/, consulté le 10.08.2020.

GILLET, F. (2019). Note de réflexion sur le numérique au cegesoma (document interne). Note de vision, CegeSoma.

GODBY, J., SMITH-YOSHIMURA, K., WASHBURN, B., DAVIS, K., DETLING, K., ESLAO, C. F., FOLSOM, S., LI, X., MCGEE, M., MILLER, K., MOODY, H., TOMREN, H. et THOMAS, C. (2019). Creating Library Linked Data with Wikibase : Lessons Learned from Project Passage. [En ligne], https://www.oclc.org/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage.html, consulté le 06.08.2019.

GOOD, B. M., BURGSTALLER-MUEHLBACHER, S., MITRAKA, E., PUTMAN, T., SU, A. I. et WAAGMEESTER, A. (2016). Opportunities and Challenges Presented by Wikidata in the Context of Biocuration. *In ICBO/BioCreative*.

GOTOVITCH, J. (2005). Éditorial. *Bulletin du CegeSoma*, (39):3–5. [En ligne], https://www.cegesoma.be/docs/media/Bulletins/bulletin39.pdf, consulté le 06.08.2020.

GOULD, T. (2018). Using the hive mind : Wikidata integration and artist pages. [En ligne], https://medium.com/nationalgalleries-digital/using-the-hive-mind-wikidata-integration-and-artist-pages-618574fa9168, consulté le 19.02.2019.

GREENE, M. et MEISSNER, D. (2005). More product, less process : Revamping traditional archival processing. *The American Archivist*, 68(2):208–263.

GROUPE AFNOR CG46/CN357/GE3 (2009). Faire un répertoire ou un inventaire simple en EAD (Description Archivistique Encodée). Manuel d'encodage, version 1.1. Manuel, Direction des Archives de France - Département de l'innovation technologique et de la normalisation. [En ligne], https://www.enssib.fr/bibliotheque-numerique/documents/62240-faire-un-repertoire-ou-un-inventaire-simple-en-ead-description-archivistique-encodee.pdf, consulté le 11.08.2020.

GUEGUEN, G., da FONSECA, V., PITTI, D. et GRIMOÜARD, C. (2013). Toward an International Conceptual Model for Archival Description : A Preliminary Report from the International Council on Archives' Experts Group on Archival Description. *The American Archivist*, 76(2):567–584.

GUILLAUD, H. (2019). De l'explicabilité des systèmes : les enjeux de l'explication des décisions automatisées. [En ligne], http://www.internetactu.net/2019/11/14/de-lexplicabilite-des-systemes-les-enjeux-de-lexplication-des-decisions-automatisees/, consulté le 01.06.2020.

GÜREL, E. et TAT, M. (2017). SWOT analysis : a theoretical review. *Journal of International Social Research*, 10(51):994–1006.

HARRISON, H., BIRKS, M., FRANKLIN, R. et MILLS, J. (2017). Case study research : Foundations and methodological orientations. *In Forum Qualitative Sozialforschung/Forum : Qualitative Social Research*, volume 18. [En ligne], http://www.qualitative-research.net/index.php/fqs/article/view/2655, consulté le 11.06.2020.

HARTIG, O. (2019). Position Statement : The RDF* and SPARQL* Approach to Annotate Statements in RDF and to Reconcile RDF and Property Graphs. [En ligne], https://blog.liu.se/olafhartig/2019/01/10/position-statement-rdf-star-and-sparql-star/, consulté le 12.12.2020.

HASSLER, M. et FLIEDL, G. (2006). Text preparation through extended tokenization. *WIT Transactions on Information and Communication Technologies*, 37.

HEBERLEIN, R. (2019). On the flipside : Wikidata for cultural heritage metadata through the example of numismatic description. *In Proceedings of IFLA WLIC 2019*. [En ligne], http://library.ifla.org/2492/, consulté le 03.06.2020.

HELLEC, F. (2014). Le rapport au terrain dans une thèse cifre. *Sociologies pratiques*, (1):101–109.

HELMREICH, A., BROSENS, K., van den HEUVEL, C., SCHELTJENS, S., van GINHOVEN, S., PUGH, E. et TRUYEN, F. (2019). Art History and Big Data : Complex Collaborations between Institutions and Researchers.

HENDA, M. B. (2018). L'ingénierie des corpus. [En ligne], https://cel.archives-ouvertes.fr/cel-01716602/, consulté le 11.06.2020.

HERBERT, D. et HOTT, R. (2019). Lettuce Into the Meal : A SNAC Update. Intervention dans le cadre de la WikidataCon 2019, [En ligne], https://docs.google.com/presentation/d/15dstNlrIxU8A4zC53BI_Omy_Ljdavy8tovbgGghD1G4, consulté le 01.09.2020.

HERNÁNDEZ, D., HOGAN, A. et KRÖTZSCH, M. (2015). Reifying RDF : What Works Well With Wikidata ? *In Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems co-located with 14th International Semantic Web Conference (ISWC 2015)*, pages 32–47.

HITZLER, P. (2020). Semantic Web : A Review Of The Field. *In Proceedings of the Semantic Web Conference 2020 (author draft)*.

HOFFMAN, R. R., KLEIN, G. et MUELLER, S. T. (2018). Explaining explanation for "explainable ai". *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201. SAGE Publications Sage CA : Los Angeles, CA.

HUNGENAERT, J. (2016). Report concerning the interviews with the staff members of the cegesoma. Rapport, CegeSoma.

HUNGENAERT, J. et GILLET, F. (2017). Studying user's digital practices and needs in Archives and Libraries. Final Report of the MADDLAIN project. Rapport, State Archives of Belgium, Centre for Historical Research and Documentation on War and Contemporary Society, Royal Library of Belgium, Département des Sciences et technologies de l'Information et de la Communication (ULB), Imec.

HWANG, KAREN LI-LUN (2017). The Vision of Linked Open Data : Martin Wong and the METRO Network. [En ligne], https://mnylc.org/fellows/2017/08/03/lod_metro/, consulté le 02.09.2020.

HYVÖNEN, E., HEINO, E., LESKINEN, P., IKKALA, E., KOHO, M., TAMPER, M., TUOMINEN, J. et MÄKELÄ, E. (2016). WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. *In European Semantic Web Conference*, pages 758–773. Springer.

INTERNATIONAL COUNCIL ON ARCHIVES (1992). Statement of Principles Regarding Archival Description. *Archivaria*, (34):8–16. [En ligne], https://archivaria.ca/index.php/archivaria/article/view/11837, consulté le 07.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (1996a). Code de déontologie des archivistes. [En ligne], https://www.ica.org/sites/default/files/ICA_1996-09-06_code%20of%20ethics_FR.pdf, consulté le 11.08.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (1996b). Isaar (cpf) : norme internationale sur les notices d'autorité archivistiques relatives aux collectivités, aux personnes et aux familles : 1ère édition adoptée par la commission ad hoc sur les normes de description, paris, france, 15-20 novembre 1995.

INTERNATIONAL COUNCIL ON ARCHIVES (2000). ISAD(G) : norme générale et internationale de description archivistique : 2e édition adoptée par le Comité sur les normes de description, Stockholm, Suède, 19-22 septembre 1999. [En ligne], https://www.ica.org/fr/ressources-publiques/normes, consulté le 11.08.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (2004). Isaar (cpf) : norme internationale sur les notices d'autorité archivistiques relatives aux collectivités, aux personnes et aux familles : 2e édition adoptée par le comité sur les normes de description, canberra, australie, 27-30 octobre 2003.

INTERNATIONAL COUNCIL ON ARCHIVES (2016a). Appel à commentaires : Publication de Records in Contexts par l'EGAD. [En ligne], https://www.ica.org/fr/appel-a-commentaires-publication-de-records-in-contexts-par-l-egad, consulté le 02.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (2016b). Au sujet du comité des normes et bonnes pratiques. [En ligne], https://www.ica.org/fr/node/14803, consulté le 02.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (2016c). Records in Contexts - Modèle conceptuel. [En ligne], https://www.ica.org/fr/records-in-contexts-modele-conceptuel, consulté le 02.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (2019). Publication de Records in Contexts-Ontology v0.1 et de la prévisualisation de Records in Contexts-Conceptuel Model v0.2. [En ligne], https://www.ica.org/fr/publication-d

e-records-in-contexts-ontology-v01-et-de-la-previsualisation-de-records-in-contexts, consulté le 02.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES (2020). Le conseil international des archives. [En ligne], https://www.ica.org/fr/le-conseil-international-des-archives, consulté le 02.09.2020.

INTERNATIONAL COUNCIL ON ARCHIVES EXPERT GROUP ON ARCHIVAL DESCRIPTION (2016). Records in Contexts. A Conceptual Model for Archival Description. Consultation Draft v0.1. [En ligne], https://www.ica.org/fr/records-in-contexts-modele-conceptuel, consulté le 11.08.2020.

INTERNATIONAL COUNCIL ON ARCHIVES EXPERT GROUP ON ARCHIVAL DESCRIPTION (2019a). International Council on Archives Records in Contexts Ontology (ICA RiC-O) version 0.1. [En ligne], https://www.ica.org/standards/RiC/ontology.html, consulté le 11.08.2020.

INTERNATIONAL COUNCIL ON ARCHIVES EXPERT GROUP ON ARCHIVAL DESCRIPTION (2019b). Records in Contexts. A Conceptual Model for Archival Description. Consultation Draft v0.2 (preview). [En ligne], https://www.ica.org/fr/records-in-contexts-modele-conceptuel, consulté le 11.08.2020.

INVENTAIRE.IO (2020). Source des données. [En ligne], https://wiki.inventaire.io/wiki/Data?lang=fr#amorces-de-données-du-web, consulté le 13.09.2020.

ISMAYILOV, A., KONTOKOSTAS, D., AUER, S., LEHMANN, J., HELLMANN, S. *et al.* (2018). Wikidata through the Eyes of DBpedia. *Semantic Web*, 9(4):493–503.

ISO (2005). 9000 : 2005. Norme, International Organization for Standardization, Genève.

JACKSON, S. J. (2014). Rethinking Repair. *Media technologies : Essays on communication, materiality, and society*, pages 221–39.

Janssens de BISTHOVEN, B. (2020). Analyse critique de l'évolution de l'EAD et de son implémentation via les logiciels AtoM et ArchivesSpace. Mémoire de D.E.A., Université libre de Bruxelles.

JESSAMY, C. (18 février 2016). Battle babies. *The National Archives (blog)*. [En ligne], https://blog.nationalarchives.gov.uk/battle-babies/, consulté le 26.08.2020.

JOHANNIC-SETA, F. (2017). Le futur FNE : vers une vraie coproduction. *Ar(abes)ques*, (85):16–17. [En ligne], http://fr.1001mags.com/parution/arabesques/numero-85-avr-mai-jun-2017, consulté le 01.04.2019.

JOHANNIC-SETA, F. et AYMONIN, D. (2019). En guise de conclusion... Intervention dans le cadre de la 4e journée professionnelle Métadonnées en bibliothèques (15 novembre 2019), [En ligne], https://www.transition-bibliographique.fr/wp-content/uploads/2019/11/10_Joannic-Seta_Aymonin.pdf, consulté le 14.09.2020.

JOHNSTON, P. (2011). Two changes to the model and some definitions. [En ligne], http://locah.archiveshub.ac.uk/2011/02/16/two-changes-to-the-model-and-some-definitions/, consulté le 30.09.2020.

JONES, T. (2018). Hello, my name is _____ : Searching for names is not always straightforward. [En ligne], https://diff.wikimedia.org/2018/05/08/searching-for-names-is-not-always-straightforward/, consulté le 27.08.2020.

JOUISON-LAFFITTE, E. (2009). La recherche action : oubliée de la recherche dans le domaine de l'entrepreneuriat. *Revue de lEntrepreneuriat*, 8(1):1–35.

JOURNAL OFFICIEL DE L'UNION EUROPÉENNE (2019). Directive (UE) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public. [En ligne], https://eur-lex.europa.eu/eli/dir/2019/1024/, consulté le 28.09.2020.

@JUERGENKETT (2018). That sounds promising. We evaluate wikibase as a possible « second home » for the GND. It would be good to share experiences and thoughts here. [Tweet], https://web.archive.org/web/20200423085318/https://twitter.com/JuergenKett/status/1022504057779511302.

KAUSHIK, A. (2018). SWOT analysis of MOOCs in library and information science domain. *Library Hi Tech News*.

KESTELOOT, C. (2019). Histoire publique - note de vision. Note de vision, CegeSoma.

KLEIN, M., GUPTA, H., RAI, V., KONIECZNY, P. et ZHU, H. (2016). Monitoring the gender gap with wikidata human gender indicators. *In Proceedings of the 12th International Symposium on Open Collaboration*, pages 1–9.

KLEIN, M. et KYRIOS, A. (2013). Viafbot and the integration of library data on wikipedia. *Code4lib journal*, (22).

KOHO, M., HEINO, E., HYVÖNEN, E. *et al.* (2016). SPARQL Faceter-Client-side Faceted Search Based on SPARQL. *In LIME/SemDev - ESWC*.

KOHO, M., IKKALA, E. et HYVÖNEN, E. (2019a). Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web. *In CEUR Workshop Proceedings*. [En ligne], https://seco.cs.aalto.fi/publications/2019/koho-et-al-reassembling-prisoner-biographies-2019.pdf, consulté le 30.09.2020.

KOHO, M., IKKALA, E., LESKINEN, P., TAMPER, M., TUOMINEN, J. et HYVÖNEN, E. (2019b). WarSampo Knowledge Graph : Finland in the Second World War as Linked Open Data. *Semantic Web–Interoperability, Usability, Applicability*.

KOLKMAN, D. (2016). Maintenance in progress ? [En ligne], https://themaintainers.org/blog/2016/6/3/maintenance-in-progress, consulté le 02.09.2020.

KRIEGER, N. et ABES (2016). Autorités personnes physiques. création dans le sudoc. [En ligne], https://wiki.scd.unistra.fr/_media/collections/entrees_traitements/catalogage_dans_le_sudoc/autorites_personnes_physiques.pdf, consulté le 27.08.2020.

LANDON, A. (2015). S'observer / Participer : la recherche en contrat CIFRE, construire une démarche de recherche-action autour du projet urbain. *In Rencontres doctorales en architecture : Quels apports entre recherche et projet dans les disciplines de l'architecture, de l'urbanisme, du paysage et du design ?, septembre 2015, Marseille, France*.

LARDINOIS, Y., TEYAR, S., BLAN, R., COENEN, A., SOYEZ, S. et Van der EYCKEN, J. (2019). State Archives Management (SAM). Manuel d'utilisation. Version 1.3. Manuel, Archives de l'État.

LE PAPE, P. (2015). Il n'y aura pas de grand soir catalographique : aujourd'hui c'est déjà demain. Intervention dans le cadre de la journée d'étude ADBU-ABES « Cataloguer demain, conduire le changement » (13 février 2015), [En ligne], https://adbu.fr/wp-content/uploads/2015/02/Ph_Le_Pape_ADBU_ABES_040215.pdf, consulté le 12.09.2020.

LEHMANN, J., ISELE, R., JAKOB, M., JENTZSCH, A., KONTOKOSTAS, D., MENDES, P. N., HELLMANN, S., MORSEY, M., VAN KLEEF, P., AUER, S. *et al.* (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

LEIBNITIANA.EU (2019). Architecture. [En ligne], https://leibnitiana.eu/architecture, consulté le 01.11.2019.

LEIGH, D. (2009). SWOT analysis. *Handbook of Improving Performance in the Workplace*, 1-3:115–140.

LEMUS-ROJA, M. et PINTSCHER, L. (2018). *Leveraging Wikipedia : Connecting Communities of Knowledge*, chapitre Wikidata and libraries : Facilitating open knowledge, pages 145–158. ALA Editions.

LESKINEN, P., KOHO, M., HEINO, E., TAMPER, M., IKKALA, E., TUOMINEN, J., MÄKELÄ, E. et HYVÖNEN, E. (2017). Modeling and using an actor ontology of second world war military units and personnel. *In International Semantic Web Conference*, pages 280–296. Springer.

LEVY, R. (2005). Les doctorants cifre : médiateurs entre laboratoires de recherche universitaires et entreprises. *Revue d'économie industrielle*, 111(1):79–96.

LI, Z., AVGERIOU, P. et LIANG, P. (2015). A systematic mapping study on technical debt and its management. *Journal of Systems and Software*, 101: 193–220.

LIEW, C. L. (2009). Digital library research 1997-2007. *Journal of Documentation*, 65(2):245–266.

LINKED DATA FINLAND (2020). Project. [En ligne], http://www.ldf.fi/project.html, consulté le 30.09.2020.

LINKING LIVES (2011a). About linking lives. [En ligne], http://linkinglives.archiveshub.ac.uk/sample-page/, consulté le 30.09.2020.

LINKING LIVES (2011b). Do not underestimate cleaning your data ! [En ligne], http://linkinglives.archiveshub.ac.uk/2012/03/08/do-not-underestimate-cleaning-your-data/, consulté le 30.09.2020.

LOVINK, G. (2008). The Society of the Query and the Googlisation of Our Lives A Tribute to Joseph Weizenbaum. *Eurozine*. [En ligne], https://www.eurozine.com/the-society-of-the-query-and-the-googlization-of-our-lives/, consulté le 09.09.2020.

LOVINS, D. et HILLMANN, D. (2017). Broken-world vocabularies. *D-Lib Magazine*.

LUYTEN, D. (2018). Een groeipad voor het fundamenteel historisch onderzoek in het cegesoma (v.1.01). Note de vision, CegeSoma.

MAERTEN, F. (2013). Fonds documentaire. Lettres d'adieu des résistants de Belgique exécutés en 1940-1944 AA2346. Note explicative. [En ligne], https://www.cegesoma.be/docs/Invent/AA_2346_Note_explicative.pdf, consulté le 30.09.2020.

MAERTEN, F. (2019). Note : Accompagnement du public dans les collections. Note de vision, CegeSoma.

MAERTEN, F. (2020). *Papy était-il un héros ? Sur les traces des hommes et des femmes dans la Résistance pendant la Seconde Guerre mondiale*. Éditions Racine, Bruxelles.

MAERTEN, F., DEBRUYNE, E. et MAERTEN (2011). En guise d'adieu. les dernières lettres des résistants de belgique exécutés lors des deux conflits mondiaux. *In Écrire sous l'occupation. Du non-consentement à la résistance : France-Belgique-Pologne 1940-1945*. [En ligne], https://books.openedition.org/pur/110987, consulté le 06.08.2020.

MALYSHEV, S., KRÖTZSCH, M., GONZÁLEZ, L., GONSIOR, J. et BIELEFELDT, A. (2018). Getting the most out of wikidata : Semantic technology usage in wikipedia's knowledge graph. *In International Semantic Web Conference*, pages 376–394. Springer.

MATTERN, S. (2018). Maintenance and Care. *Places Journal*.

MATTHEW, R. (2012). The Wikipedia data revolution. [En ligne], https://blog.wikimedia.org/2012/03/30/the-wikipedia-data-revolution/, consulté le 14.05.2020.

MCKENZIE, P. (2010). Falsehoods Programmers Believe About Names. [En ligne], https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/, consulté le 27.08.2020.

MEDIAWIKI (2017). Talk :Wikibase/Installation - Installing the "right" version. [En ligne], https://www.mediawiki.org/wiki/Manual_talk:Managing_data_in_MediaWiki, consulté le 22.05.2020.

MEDIAWIKI (2019). Manual talk :Managing data in MediaWiki - Is Wikibase installation 'easy' ? [En ligne], https://www.mediawiki.org/wiki/Manual_talk:Managing_data_in_MediaWiki, consulté le 22.05.2020.

MEDIAWIKI (2020a). Cycle de vie des versions. [En ligne], https://www.mediawiki.org/wiki/Version_lifecycle/fr, consulté le 28.09.2020.

MEDIAWIKI (2020b). Phabricator/Help. [En ligne], https://www.mediawiki.org/wiki/Phabricator/Help#Statistics, consulté le 11.08.2020.

MEDIAWIKI (2020). Wikibase/DataModel/Primer. [En ligne], https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer, consulté le 22.05.2020.

MEEMOO (2019). Publishing and sourcing linked data : where do you start ? [En ligne], https://meemoo.be/en/publications/publishing-and-sourcing-linked-data-where-do-you-start, consulté le 08.07.2020.

MEISSNER, D. et GREENE, M. A. (2010). More Application while Less Appreciation : The Adopters and Antagonists of MPLP. *Journal of Archival Organization*, 8(3-4):174–226.

MILLER, M. (2018). Wikibase for Research Infrastructure — Part 1. [En ligne], https://medium.com/@thisismattmiller/wikibase-for-research-infrastructure-part-1-d3f640dfad34, consulté le 21.04.2020.

MINISTÈRE DE LA CULTURE (2016). Renkan, un outil d'appropriation des données. [En ligne], https://www.culture.gouv.fr/Sites-thematiques/Innovation-numerique/Donnees-publiques/Renkan-un-outil-d-appropriation-des-donnees, consulté le 02.09.2020.

MOIRAND, S. (2004). L'impossible clôture des corpus médiatiques La mise au jour des observables entre catégorisation et contextualisation. *Tranel*, (40):71–92.

MOLINIÉ, C. (2020). Les vases communicants : la collaboration entre musées et projets Wikimedia pour le partage du patrimoine culturel. [En ligne], https://www.wikimedia.fr/collaboration-musees-wikimedia-partage-patrimoine-culturel/, consulté le 20.05.2020.

MONITEUR BELGE (1955). Loi du 24 juin 1955 relative aux archives. [En ligne], http://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=fr&la=F&cn=1955062430&table_name=loi, consulté le 10.08.2020.

MONITEUR BELGE (2018a). Loi du 21 décembre 2018 portant des dispositions diverses en matière de justice. [En ligne], https://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=fr&la=F&table_name=loi&cn=2018122109, consulté le 23.09.2020.

MONITEUR BELGE (2018b). Loi du 30 juillet 2018 relative à protection des personnes physiques à l'égard des traitements de données à caractère personnel. [En ligne], https://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=fr&la=F&table_name=loi&cn=2018073046, consulté le 10.08.2020.

MÜLLRICK, S. (2019). LearningWikibase. [En ligne], https://github.com/samu-workopen/learningwikibase#readme, consulté le 23.05.2020.

NEUBERT, J. (2017). Wikidata as a Linking Hub for Knowledge Organization Systems ? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings. *In Proceedings of the 17th European Networked Knowledge Organization Systems Workshop*, pages 14–25. [En ligne], http://ceur-ws.org/Vol-1937/paper2.pdf, consulté le 19.07.2019.

NIU, J. (2016). Linked data for archives. *Archivaria*, 82(1):83–110.

NOUVELLET, A., D'ALCHÉ-BUC, F., BEAUDOUIN, V., PRIEUR, C. et ROUEFF, F. (2017). Discovery of usage patterns in digital library web logs using markov modeling. *In Proceeding of KDD'17 (à venir)*, Halifax, Canada.

NOY, N., GAO, Y., JAIN, A., NARAYANAN, A., PATTERSON, A. et TAYLOR, J. (2019). Industry-scale knowledge graphs : lessons and challenges. *Queue*, 17(2):48–75.

NUGROHO, A., VISSER, J. et KUIPERS, T. (2011). An empirical model of technical debt and interest. *In Proceedings of the 2nd workshop on managing technical debt*, pages 1–8.

ODIS (2020). De ideale ODIS-steekkaart ? : uitgangspunten en tips. [En ligne], https://www.odis.be/hercules/docs/Ontmoetingsdag_2019_workshop_3.pdf, consulté le 06.08.2020.

OHLIG, J. (2018). Gemeinsam wieder Neuland betreten : Die Deutsche Nationalbibliothek und Wikimedia Deutschland. [En ligne], https://blog.wikimedia.de/2018/11/02/gemeinsam-wieder-neuland-betreten-die-deutsche-nationalbibliothek-und-wikimedia-deutschland/, consulté le 21.04.2020.

on Archives Expert Group on ARCHIVAL DESCRIPTION, I. C. (2020). Ica egad records in contexts-ontology (ric-o) github repository web pages - about. [En ligne], https://ica-egad.github.io/RiC-O/about.html, consulté le 13.12.2020.

ORR, J. E. (1996). *Talking about Machines : An Ethnography of a Modern Job*. Cornell University Press.

ORR, J. E. (2006). Ten years of talking about machines. *Organization Studies*, 27(12):1805–1820.

OTTOSSON, S. (2003). Participation action research- : A key to improved knowledge of management. *Technovation*, 23(2):87–94.

PALMER, J. W. (1986). Subject authority control and syndetic structure-myth and realities : An inquiry into certain subject heading practices and some questions about their implications. *Cataloging & classification quarterly*, 7(2):71–95.

PATRICK, L., VAN DOORSALER, R. et VELLE, K. (2014). Culture fédérale : La mémoire du pays est-elle en péril ? *La Libre.be*. [En ligne], https://www.lalibre.be/debats/opinions/culture-federale-la-memoire-du-pays-est-elle-en-peril-544e46cd3570a5ad0ede5fb7, consulté le 06.08.2020.

PAUL, S. (2017). Virtual research environments. Rapport, MADDLAIN Project. [En ligne], https://www.cegesoma.be/docs/images/stories/ceges/Projets_en_cours/MADDLAIN_VRE_ENG_61p.pdf, consulté le 19.07.2020.

PAULOZZI, L. J., COX, C. S., WILLIAMS, D. D. et NOLTE, K. B. (2008). John and jane doe : the epidemiology of unidentified decedents. *Journal of forensic sciences*, 53(4):922–927.

PELLISSIER TANON, T., VRANDEČIĆ, D., SCHAFFERT, S., STEINER, T. et PINTSCHER, L. (2016). From freebase to wikidata : The great migration. *In Proceedings of the 25th international conference on world wide web*, pages 1419–1428.

PERSONALDATA.IO (2020). PersonalData.IO Community Call 05 : technical aspects of our Wikibase. [Vidéo], https://invidio.us/watch?v=D950_JoNDPI, consulté le 01.07.2020.

PINTSCHER, L. (2017). User :Lydia Pintscher (WMDE)/CC-0. [En ligne], https://www.wikidata.org/wiki/User:Lydia_Pintscher_(WMDE)/CC-0, consulté le 14.05.2020.

PINTSCHER, L. (2018). Wikidata : Sixth Birthday - Message from dev team. [En ligne], https://www.wikidata.org/wiki/Wikidata:Sixth_Birthday/Message_from_dev_team, consulté le 21.04.2020.

PINTSCHER, L. (2020). Wikibase Ecosystem - taking Wikidata further. Intervention dans le cadre de l'édition 2020 du FOSDEM 2020, [En ligne], https://archive.fosdem.org/2020/schedule/event/wikibase_ecosystem/, consulté le 14.09.2020.

PINTSCHER, L. et OHLIG, J. (2019). Everyone gets one - Wikibase and the Wikibase Ecosystem. Intervention dans le cadre de la Conférence Wikimania 2019, [En ligne], https://commons.wikimedia.org/w/index.php?title=File:Wikimania_2019_Everyone_gets_one_-_Wikibase_and_the_Wikibase_Ecosystem.pdf, consulté le 14.09.2020.

PINTSCHER, L., VOGET, L., KOEPPEN, M., ALEYNIKOVA, E., MANICKI, L., DITTRICH, J., SHUTY, R., OHLIG, J., MÜLLER, B., BITTAKER, A., ISLER, R., MINOR, J. et VERSHBOW, B. a. (2019a). Vision and high level overview for Wikidata and Wikibase. [En ligne], https://meta.wikimedia.org/wiki/File:Vision_and_high_level_overview_for_Wikidata_and_Wikibase.pdf, consulté le 21.04.2020.

PINTSCHER, L., VOGET, L., KOEPPEN, M., ALEYNIKOVA, E., MANICKI, L., DITTRICH, J., SHUTY, R., OHLIG, J., MÜLLER, B., ISLER, R., MINOR, J., VERSHBOW, B. et BITTAKER, A. (2019b). Strategy for the Wikibase Ecosystem. [En ligne], https://meta.wikimedia.org/wiki/File:Strategy_for_Wikibase_Ecosystem.pdf, consulté le 21.04.2020.

PITTI, D., HU, R., LARSON, R., TINGLE, B. et TURNER, A. (2015). Social Networks and Archival Context : From Project to Cooperative Archival Program. *Journal of Archival Organization*, 12(1-2):77–97.

PITTI, D., STOCKTING, B. et CLAVAUD, F. (2018). An introduction to "Records in Contexts" : an archival description draft standard. *Comma*, 2016(1-2):173–189.

PLANE, J.-M. (1998). Pour une approche ethnométhodologique de la PME. *Revue internationale PME Économie et gestion de la petite et moyenne entreprise*, 11(1):123–140.

PLANTIN, J.-C. (2018). Data cleaners for pristine datasets : Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, 44(1):52–73.

@PODEHAYE (2019). L'écosystème d'outils qu'il est possible de constituer au sein de Wikibase permet d'aller bien bien au delà du cadre strict des archives ! [Tweet], https://web.archive.org/web/20190918133400/https://twitter.com/podehaye/status/1174310064276070405.

POLITIQUE SCIENTIFIQUE FÉDÉRALE (2020). BRAIN-be. [En ligne], https://www.belspo.be/belspo/brain-be/index_fr.stm, consulté le 02.09.2020.

POPOVICI, B. (2019). Records in contexts : vers un nouveau niveau dans la description archivistique ? *Archives*, 48(2):7–39.

POUCHOL, J. (2016). *Mutualiser les pratiques documentaires. Bibliothèques en réseau*. ENSSIB.

POULTER, M. (2019). Teaching SPARQL as a Foreign Language. Intervention dans le cadre de la WikidataCon 2017, [En ligne], https://commons.wiki

media.org/wiki/File:Martin_Poulter_2019_Teaching_SPARQL_as_a_Foreign_Language.pdf, consulté le 30.09.2020.

POUPEAU, G. (2019). Why i don't use semantic web technologies anymore, even if they still influence me ? [En ligne], http://www.lespetitescases.net/why-I-dont-use-semantic-web-technologies-anymore-even-if-they-still-influence-me, consulté le 01.06.2020.

PROFFITT, M. (2019). Exploration of Wikibase by Librarians. Intervention dans le cadre de la Conférence Wikimania 2019, [En ligne], https://web.archive.org/web/20200914180007if_/https://docs.google.com/presentation/d/1CfXqo2Awltqc4576uEofTkzvKySN5apjJHndYM3YuKg/edit#slide=id.g5f3115dd98_2_85, consulté le 14.09.2020.

RAGNARSDÓTTIR, S. K. (21 juin 2019). Stúlkur mega nú heita Ari og drengir Anna. *RÚV*. [En ligne], https://www.ruv.is/frett/stulkur-mega-nu-heita-ari-og-drengir-anna, consulté le 26.08.2020.

RANADE, S. (2015). Traces Through Time. Intervention dans le cadre du Digital History Seminar/Archives and Society Seminar (23 juin 2015), [En ligne], https://fr.slideshare.net/historyspot/ranade-ihr-june-2015, consulté le 30.09.2020.

RANADE, S. (2016a). Making connections : tracing people through our collection. [En ligne], https://blog.nationalarchives.gov.uk/making-connections-tracing-people-collection/, consulté le 02.09.2020.

RANADE, S. (2016b). Traces through time : A probabilistic approach to connected archival data. *In 2016 IEEE International Conference on Big Data (Big Data)*, pages 3260–3265. IEEE.

RAO, D., MCNAMEE, P. et DREDZE, M. (2013). Entity Liking : Finding Extracted Entities in a Knowledge Base. *In Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

RAPOPORT, R. N. (1970). Three dilemmas in action research : with special reference to the tavistock experience. *Human relations*, 23(6):499–513.

RAVALET, E., LUCAS, J.-F. et LOHOU, A. (2017). Les retours d'une exploration méthodologique croisant données Twitter, recrutement via Facebook et questionnaires web. La mobilité des jeunes dans et par les réseaux sociaux. [En ligne], http://journals.openedition.org/netcom/2734, consulté le 10.06.2020.

REGESTA.COM (2015). Lodlam 2015. regesta vince il grand prize con open memory project. [En ligne], https://www.regesta.com/2015/06/30/lodla

m-2015-regesta-vince-il-gran-prize-con-open-memory-project, consulté le 30.09.2020.

REY-BELLET, G. (2015). Le wikipédien en résidence : médiateur entre les institutions culturelles et la communauté wikimédia. *Arbido*, (3/15):11–12.

RIBES, D. et FINHOLT, T. A. (2009). The Long Now of Infrastructure : Articulating Tensions in Development. *Journal of the Association for Information Systems (JAIS)*.

RIGBY, J., COX, D. et JULIAN, K. (2018). Journal peer review : a bar or bridge ? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics*, 114(3):1087–1105.

RIZZA, E., CHARDONNENS, A. et van HOOLAND, S. (2019). Close-reading of Linked Data : a case study in regards to the quality of online authority files. *arXiv preprint arXiv :1902.02140*.

ROSEN, R. J. (2012). Teaching Wikipedia to Write Itself. [En ligne], https://www.theatlantic.com/technology/archive/2012/04/teaching-wikipedia-to-write-itself/255363/, consulté le 14.05.2020.

ROUCHI, C. (2017). Réflexivité et recherche-action en contrat CIFRE, quand les contraintes du terrain deviennent opportunités. *Nouvelles perspectives en sciences sociales*, 13(1):211–224.

ROUX, M. et POVEDA, R. (2019). État des lieux international et perspectives sur la visualisation des données. Intervention dans le cadre de la 4e journée professionnelle Métadonnées en bibliothèques (15 novembre 2019), [En ligne], https://www.transition-bibliographique.fr/wp-content/uploads/2019/11/07_Roux_Poveda.pdf, consulté le 12.09.2020.

ROY, M. et PRÉVOST, P. (2013). La recherche-action : origines, caractéristiques et implications de son utilisation dans les sciences de la gestion. *Recherches qualitatives*, 32(2):129–151.

RUSSELL, A. et VINSEL, L. (2016). Hail the maintainers. *Aeon*. [En ligne], https://aeon.co/essays/innovation-is-overvalued-maintenance-often-matters-more, consulté le 30.09.2020.

RUSSELL, A. L. et VINSEL, L. (2018). After Innovation, Turn to Maintenance. *Technology and Culture*, 59(1):1–25.

SÄLGÖ, M. (2020). Carl Larsson who is that - sadly Europeana doesnt know –> #Metadatadebt. [En ligne], http://minancestry.blogspot.com/2020/03/carl-larsson-who-is-that-sadly.html, consulté le 31.08.2020.

@SALGO60 (2019). I see we have bad curated archives delivering RDF but use « strings[,] not things » and we have to pay a prize of not finding what we are looking for I see a big debt in *aggregators* like #Europeana where you *lose* so much #me[t]adata that it feels useless. [Tweet], https://web.archive.org/web/20191028115448/https://twitter.com/salgo60/status/1188764613212549120.

SALLANTIN, T. (2012). Comment traduire « sustainable development » ? [En ligne], http://netoyens.info/index.php/contrib/09/06/2012/comment-traduire-sustainable-development, consulté le 02.09.2020.

SAVOJA, M. et VITALI, S. (2008). Authority control for creators in italy : theory and practice. *Journal of Archival Organization*, 5(1-2):121–147.

SCALLA, A. (2017). Les méthodes agiles en bibliothèque. Mémoire de D.E.A., Université de Lyon.

@SCHOLL_I (2018). Das war auf der GNDCon heute wirklich beeindruckend : #Wikibase in aller Munde. @JuergenKett von der @DNB_Aktuelles am Ende der Keynotes : « Wikibase ist als Heilsbringer oft genug genannt worden. Da reihe ich mich jetzt einfach ein. ». [Tweet], https://web.archive.org/web/20200423085721/https:/twitter.com/scholl_i/status/1069642587974393856.

SCHUMAN, L. A. (1987). *Plans and situated actions : The problem of human-machine communication*. Cambridge University Press.

SCHÜTZE, H., MANNING, C. D. et RAGHAVAN, P. (2008). *Introduction to Information Retrieval*, volume 39. Cambridge University Press Cambridge.

SCOTT, D. et ALLISON-CASSIN, S. (2018). Wikibase : configure, customize, and collaborate. Loading data from Wikidata. [En ligne], https://stuff.coffeecode.net/2018/wikibase-workshop-swib18.html#_loading_data_from_wikidata, consulté le 11.05.2020.

SELENER, J. D. (1997). *Participatory action research and social change : Approaches and critique.* Cornell Participatory Action Research Network, Cornell University.

SENALADA, T. (2019). Wikibase & NOEMI. Proof of Concept about Metadata production software for the French national Library. Intervention dans le cadre de la WikidataCon 2019, [En ligne], https://commons.wikimedia.org/w/index.php?title=File:Wikibase_for_FNE.pdf, consulté le 15.05.2020.

SHORLAND, A. (2017). Wikibase docker images. [En ligne], https://addshore.com/2017/12/wikibase-docker-images/, consulté le 21.04.2020.

SHORLAND, A. (2019a). An introduction to WBStack. [En ligne], https://addshore.com/2019/11/an-introduction-to-wbstack/, consulté le 21.04.2020.

SHORLAND, A. (2019b). wikibase-docker, Mediawiki Wikibase update. [En ligne], https://addshore.com/2019/01/wikibase-docker-mediawiki-wikibase-update/, consulté le 22.05.2020.

SHORLAND, A. (2019c). Wikidata Architecture Overview (diagrams). [En ligne], https://addshore.com/2018/12/wikidata-architecture-overview-diagrams/, consulté le 28.05.2020.

SHORLAND, A. (2020a). WBStack 2020 Update 1. [En ligne], https://addshore.com/2020/04/wbstack-2020-update-1/, consulté le 22.05.2020.

SHORLAND, A. (2020b). WBStack 2020 Update 2 (May). [En ligne], https://addshore.com/2020/05/wbstack-2020-update-2/, consulté le 22.05.2020.

SIBILLE, C. (2012a). Les normes internationales de description archivistique : origines, développements, perspectives. *Gazette des archives*, 228(4):77–90.

SIBILLE, C. (2012b). Élaborer des normes de description... et les confronter à la pratique d'aujourd'hui. *Gazette des archives*, 226(2):165–177.

SIBILLE, C. (2014). Le groupe d'experts sur la description archivistique (egad). *Flash*, (28):6. [En ligne], https://www.ica.org/sites/default/files/Flash-28-fr.pdf, consulté le 02.09.2020.

SIBILLE, C. (2017). D'hier à aujourd'hui : les évolutions de la description archivistique. *Gazette des archives*, 247(3):117–123.

SMITH-YOSHIMURA, K. (2018a). Are distributed models for vocabular maintenance viable. [En ligne], http://hangingtogether.org/?p=6672, consulté le 01.04.2019.

SMITH-YOSHIMURA, K. (2018b). The rise of wikidata as a linked data source. [En ligne], https://hangingtogether.org/?p=6775, consulté le 19.07.2019.

@SMYLES (2019a). [I] have tried to adapt the concept of *technical debt* [...] into *metadata debt* : we've cut some corners with our metadata, in order to get things done quick-n-dirty. but now we've built up metadata debt, which means we need to pay interest on it, or pay it off. [Tweet], https://web.archive.org/web/20200303104419/https://twitter.com/smyles/status/1159184295350689793.

@SMYLES (2019b). yes, exactly ! money and time. we didn't make a "mistake" before, we made a trade-off (we took on debt to go faster). and we can put off paying down the debt if we want, but there is a cost (we keep paying more-n-more interest). [Tweet], https://web.archive.org/web/20200427163354/https://twitter.com/smyles/status/1159192983742427136.

SNAC (2017). Research Use. [En ligne], https://web.archive.org/web/20170822211620/http://socialarchive.iath.virginia.edu/research_use.html, consulté le 01.09.2020.

SNAC (2020a). Becoming a SNAC Cooperative Member. [En ligne], https://portal.snaccooperative.org/node/483, consulté le 01.09.2020.

SNAC (2020b). History of SNAC. [En ligne], https://portal.snaccooperative.org/node/356, consulté le 01.09.2020.

SNAC (2020c). Research and Development (2010-2015). [En ligne], https://portal.snaccooperative.org/research_and_development, consulté le 06.10.2020.

SNAC (2020d). SNAC (Social Networks and Archival Context) August 2020 Newsletter. [En ligne], https://portal.snaccooperative.org/system/files/media/documents/Public/Communications_email_2020-08-03.pdf, consulté le 01.09.2020.

SOCIETY OF AMERICAN ARCHIVISTS (2004). Description Archivistique Encodée. Dictionnaire des balises. Traduit de l'anglais par le groupe AFNOR CG46/CN357/GE3. [En ligne], https://francearchives.fr/file/0def64f5a10f3f1ae03fdea59399a3e0755ef157/static_1066.pdf, consulté le 11.08.2020.

SOUCHIER, E. (1996). L'écrit d'écran, pratiques d'écriture & informatique. *Communication & langages*, 107(1):105–119.

SOULÉ, B. (2007). Observation participante ou participation observante ? Usages et justifications de la notion de participation observante en sciences sociales. *Recherches qualitatives*, 27(1):127–140.

SPITZ, A., DIXIT, V., RICHTER, L., GERTZ, M. et GEISS, J. (2016). State of the Union : A Data Consumer's Perspective on Wikidata and its Properties for the Classification and Resolution of Entities. *In Tenth International AAAI Conference on Web and Social Media*.

STATBEL (2018). Modification des codes INS des communes et des arrondissements administratifs à partir du 1er janvier 2019. [En ligne],

https://statbel.fgov.be/fr/nouvelles/modification-des-codes-ins-des-communes-et-des-arrondissements-administratifs-partir-du, consulté le 02.09.2020.

STATBEL (2019). Découpages géographiques. [En ligne], https://statbel.fgov.be/fr/propos-de-statbel/methodologie/classifications/geographie, consulté le 02.09.2020.

STEINER, T. (2016). Wikipedia tools for google spreadsheets. *In Proceedings of the 25th International Conference Companion on World Wide Web*, pages 997–1000.

STEVENSON, A. (2011). Final product post : Archives hub ead to rdf xslt stylesheet. [En ligne], http://locah.archiveshub.ac.uk/2011/07/01/final-product-post-ead2rdf/, consulté le 30.09.2020.

STEVENSON, J. (2012a). Linking Lives : Creating an End-User Interface Using Linked Data. *Information Standards Quarterly*, 24(2/3):14–23. [En ligne], https://www.niso.org/niso-io/2012/06/linking-lives, consulté le 30.09.2020.

STEVENSON, J. (2012b). Linking Lives : Linked Data interface. [En ligne], https://www.slideshare.net/JaneStevenson/ica2012-linkinglives, consulté le 30.09.2020.

STIBBE, H. L. (1998). Standardising description : the experience of using ISAD (G). *Lligall*, (12):132–151.

STRAUSS, A. (1988). The articulation of project work : An organizational process. *Sociological Quarterly*, 29(2):163–178.

STYVEN, D. (2020). *META*, (4):10–15.

TAPLEY HOYT, C. (2020). Ooh Na Na, What's My Name ? [En ligne], https://cthoyt.com/2020/04/18/ooh-na-na.html, consulté le 20.08.2020.

TEMMERMAN, P. (2007). Pallas a dix 10 ans. *Bulletin du CegeSoma*, (40):24–25. [En ligne], https://www.cegesoma.be/docs/media/Bulletins/Bulletin40.pdf, consulté le 06.08.2020.

TEMMERMANN, P. et WAEYENBERG, S. (2000). *Thesaurus*. Centre d'Études et de Documentation Guerre et Sociétés contemporaines (C.E.G.E.S), Bruxelles.

THIBODEAU, S. (1995). Archival Context as Archival Authority Record : The ISAAR (CPF). *Archivaria*, 40.

THOMAS, C. M. (1984). Authority control in manual versus online catalogs : an examination of see references. *Information technology and libraries*, 3(4):393–398.

TILLMAN, R. K. (2016). Opportunities for Encoding EAD for Linked Data Extraction and Publication. *Journal of Archival Organization*, 13(1-2):19–36.

TIMMS, K. (2017). Records in Contexts Conceptual Model (RiC-CM) - Consultation Draft v0.1. Overview of the Feedback Received from the Archival Community. Intervention dans le cadre de la rencontre annuelle de l'Experts Group on Archival Description, [En ligne], https://www.icar.beniculturali.it/fileadmin/risorse/Materiali_RIC_26.10.2017/Katherine_Timms_26.10.2017.pdf, consulté le 12.12.2020.

TOM, E., AURUM, A. et VIDGEN, R. (2013). An exploration of technical debt. *Journal of Systems and Software*, 86(6):1498–1516.

van HOOLAND, S., RODRÍGUEZ, E. M. et BOYDENS, I. (2011). Between commodification and engagement : On the double-edged impact of user-generated metadata within the cultural heritage sector. *Library Trends*, 59(4). [En ligne], http://hdl.handle.net/2142/26432, consulté le 11.04.2015, p. 707-720.

van HOOLAND, S. et VERBORGH, R. (2014). *Linked Data for Libraries, Archives and Museums : How to Clean, Link and Publish your Metadata*. Amer Library Assn Editions.

van VEEN, T. (2017). Wikidata as universal library thesaurus. Intervention dans le cadre de la WikidataCon 2017, [En ligne], https://commons.wikimedia.org/wiki/File:Wikidata_as_universal_library_thesaurus_-_tvv.pdf, consulté le 03.06.2020.

van VEEN, T. (2019). Wikidata : From "an" Identifier to "the" Identifier. *Information Technology and Libraries*, 38(2):72–81.

VANNESTE, W. (2013). Privacy beperkt openbaarheidcase felixarchief. Intervention dans le cadre de la journée d'étude VVBAD « Digital Archief & Privacy » (14 mai 2013), [En ligne], https://www.slideshare.net/VVBAD/privacy-felix-archief20130514, consulté le 12.09.2020.

VERBORGH, R., VAN HOOLAND, S., COPE, A. S., CHAN, S., MANNENS, E. et Van de WALLE, R. (2015). The fallacy of the multi-API culture. *Journal of Documentation*.

VOORBIJ, H. (2010). The use of web statistics in cultural heritage institutions. *Performance Measurement and Metrics*, 11(3):266–279.

VRANDEČIĆ, D. et KRÖTZSCH, M. (2014). Wikidata : a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

W3C (2014). Best Practices for Publishing Linked Data. [En ligne], https://www.w3.org/TR/ld-bp/, consulté le 27.08.2020.

W3C et ISHIDA, R. (2011). Personal names around the world. [En ligne], https://www.w3.org/International/questions/qa-personal-names.en, consulté le 27.08.2020.

WAAGMEESTER, A. (2019). Introducing Gene Wiki, Wikidata and Wikibase Tokyo : September 28th, 2019. [En ligne], https://docs.google.com/presentation/d/1IjB1lrJ_OKxPzoM3jlFThKB8Km0twhY8ocQlDdaN7b0/edit, consulté le 31.05.2019.

WAAGMEESTER, A., ESPENSCHIED, D., SHORLAND, A. et MOULDS, L. (2018a). A federated landscape of Wikibase instances, with Wikidata as a central hub. [En ligne], https://web.archive.org/web/20200514082332/https://docs.google.com/document/d/1dAa-WpWppMb4q71FgiUgZVPkp9EyMAYhurLLjZQ8qTk/edit, consulté le 14.05.2020.

WAAGMEESTER, A., THORNTON, K. et SHORLAND, A. (2018b). Using OpenStack to run a custom Wikibase. [En ligne], https://fuga.cloud/labs/using-openstack-to-run-custom-wikibase/, consulté le 14.05.2020.

WAAGMEESTER, A., WILLIGHAGEN, E. L., SU, A. I., KUTMON, M., GAYO, J. E. L., FERNÁNDEZ-ÁLVAREZ, D., GROOM, Q., SCHAAP, P. J., VERHAGEN, L. M. et KOEHORST, J. J. (2020). A protocol for adding knowledge to wikidata, a case report. *BioRxiv*.

WAIBEL, G. et ERWAY, R. (2009). Think globally, act locally : library, archive, and museum collaboration. *Museum Management and Curatorship*, 24(4):323–335.

WALLIS, R. (2019). Something For Archives in Schema.org. [En ligne], https://www.dataliberate.com/2019/04/03/something-for-archives-in-schema-org/, consulté le 12.02.2020.

WIKIBASE COMMUNITY (2018). Wikibase Community User Group Reports - 2018. [En ligne], https://meta.wikimedia.org/wiki/Wikibase_Community_User_Group/Reports/2018, consulté le 21.04.2020.

WIKIBASE COMMUNITY (2019). Wikibase Community User Group Reports - 2019. [En ligne], https://meta.wikimedia.org/wiki/Wikibase_Community_User_Group/Reports/2019, consulté le 21.04.2020.

WIKIBASE COMMUNITY (2020). Online Meeting, February 2020 - Notes. [En ligne], https://etherpad.wikimedia.org/p/WBUG_February_2020, consulté le 23.05.2020.

WIKIDATA (2018). Wikidata :WikiProject Wikidata for research/Meetups/2018-04-23-25-Antwerpen. [En ligne], https://www.wikidata.org/wiki/Wikidata:WikiProject_Wikidata_for_research/Meetups/2018-04-23-25-Antwerpen, consulté le 15.08.2020.

WIKIDATA (2019). Wikidata : Service de requête SPARQL. [En ligne], https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/fr, consulté le 17.07.2020.

WIKIDATA (2020). Portail des contraintes de propriété. [En ligne], https://www.wikidata.org/wiki/Help:Property_constraints_portal/fr, consulté le 15.09.2020.

WIKIDATA (2020a). Wikidata : Accès aux données. [En ligne], https://www.wikidata.org/wiki/Wikidata:Data_access/fr, consulté le 17.07.2020.

WIKIDATA (2020b). Wikidata : Development plan. [En ligne], https://www.wikidata.org/w/index.php?title=Wikidata:Development_plan&oldid=1151414424, consulté le 23.05.2020.

WIKIDATA (2020c). Wikidata : Glossaire. [En ligne], https://www.wikidata.org/wiki/Wikidata:Glossary/fr, consulté le 17.07.2020.

WIKIDATA (2020d). Wikidata : Identifier migration. [En ligne], https://www.wikidata.org/wiki/Wikidata:Identifier_migration, consulté le 13.05.2020.

WIKIDATA (2020e). Wikidata : Notoriété. [En ligne], https://www.wikidata.org/wiki/Wikidata:Notability/fr, consulté le 13.05.2020.

WIKIDATA (2020). Wikidata : Patrol. [En ligne], https://www.wikidata.org/wiki/Wikidata:Patrol, consulté le 15.09.2020.

WIKIDATA (2020). Wikidata : Project chat/Archive/2019/08 - Modeling the Survey of Scottish Witchcraft database on Wikidata - which property proposals to go forward with. [En ligne], https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2019/08#Modeling_the_Survey_of_Scottish_Witchcraft_database_on_Wikidata_-_which_property_proposals_to_go_forward_with, consulté le 13.05.2020.

WIKIDATA (2020a). Wikidata : Property proposal/Archive/39 : short author name. [En ligne], https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/39#short_author_name, consulté le 15.08.2020.

WIKIDATA (2020b). Wikidata : Statistiques. [En ligne], https://www.wikidata.org/wiki/Wikidata:Statistics/fr, consulté le 11.08.2020.

WIKIDATA (2020). Wikidata talk :Vandalism. [En ligne], https://www.wikidata.org/wiki/Wikidata_talk:Vandalism#Editing_rights, consulté le 27.04.2020.

WIKIMEDIA (2018). Strategy/Wikimedia movement/2017/Direction. [En ligne], https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction, consulté le 12.09.2020.

WIKIMEDIA (2020). Founding principles. [En ligne], https://meta.wikimedia.org/wiki/Founding_principles, consulté le 27.04.2020.

WIKIMEDIA DEUTSCHLAND (2019a). Wikimedia Deutschland : Zukunftprozess 2018 - Strategien. [En ligne], https://meta.wikimedia.org/wiki/Wikimedia_Deutschland/Zukunftsprozess/Strategien/EN, consulté le 21.04.2020.

WIKIMEDIA DEUTSCHLAND (2019b). Wikimedia Deutschland/Planung 2020/Ziele und Erfolgskriterien/en. [En ligne], https://meta.wikimedia.org/wiki/Wikimedia_Deutschland/Planung_2020/Ziele_und_Erfolgskriterien/en, consulté le 20.05.2020.

WIKIMEDIA DEUTSCHLAND (2020). Full Stack Developer (m/f/d). [En ligne], https://wikimedia-deutschland.softgarden.io/job/4361894/Full-Stack-Developer-m-f-d-?jobDbPVId=17933060&l=de, consulté le 22.05.2020.

WIKIMEDIA NEDERLAND (2015). GLAM-WIKI 2015 - 12 April - Training GLAMs. [Vidéo], https://youtu.be/mEEgW8rWgVQ?t=978, consulté le 20.05.2020.

WIKIPEDIA (2019). Wikipedia :Wikidata/2018 Infobox RfC. [En ligne], https://en.wikipedia.org/wiki/Wikipedia:Wikidata/2018_Infobox_RfC#Discussion, consulté le 04.06.2020.

WIKIPEDIA (2020). Rhizome (organization). [En ligne], https://en.wikipedia.org/wiki/Rhizome_(organization), consulté le 23.04.2020.

WIKIPÉDIA (2017). Discussion Wikipédia : Prise de décision/Utilisation des données Wikidata dans les articles/Archives 2. [En ligne], https://fr.wikipedia.org/wiki/Discussion_Wikipédia:Prise_de_décision/Utilisation_des_données_Wikidata_dans_les_articles/Archives_2, consulté le 04.06.2020.

WIKIPÉDIA (2019). Aide : Infobox. [En ligne], https://fr.wikipedia.org/wiki/Aide:Infobox, consulté le 14.09.2020.

WIKIPÉDIA (2020a). Le Bistrot/13 septembre 2018 : pourquoi peut-on éditer sans être inscrit ? [En ligne], https://fr.wikipedia.org/wiki/Wikipédia:Le_Bistro/13_septembre_2008#pourquoi_peut-on_éditer_sans_être_inscrit?, consulté le 27.04.2020.

WIKIPÉDIA (2020b). Projet :wikipédia/listepb/ip. [En ligne], https://fr.wikipedia.org/wiki/Projet:Wikipédia/listepb/IP, consulté le 27.04.2020.

WIKTIONNAIRE (2020). Sémantisation. [En ligne], https://fr.wiktionary.org/wiki/sémantisation, consulté le 02.09.2020.

WILTON, P. (2018). Wikidata - Q41483. Thoughts on working with Wikidata from practical experience - its benefits and drawbacks. [En ligne], https://datalanguage.com/news/wikidata-q41483, consulté le 02.06.2020.

YIN, R. K. (1984). *Case Study Research : Design and Methods*. Applied Social Research Methods Series. Sage Publications.

YOAKIM, W. (2019). Wikipédia, wikimedia commons et wikisource, un eldorado de visibilité. *Archives*, 48(2):41–81.

ZAVALINA, O. L. et ZAVALIN, V. (2018). Evaluation of metadata change in authority data over time : An effect of a standard evolution. *Proceedings of the Association for Information Science and Technology*, 55(1):593–597.

ZHOU, L., SHIMIZU, C., HITZLER, P., SHEILL, A. M., ESTRECHA, S. G., FOLEY, C., TARR, D. et REHBERGER, D. (2020). The Enslaved Dataset : A Real-world Complex Ontology Alignment Benchmark using Wikibase. *In Proceedings of the 29th ACM International Conference on Information and Knowledge Managamant (CIKM '20)*. ACM New York, NY, USA. [En ligne], https://daselab.cs.ksu.edu/sites/default/files/2020-CIKM-enslaved-alignment.pdf, consulté le 19.08.2020.

ZUIJDAM, F., HELMUS, W., PETER, V., ROMAN, L., SCHIPPER, J., TERRIER, A. et van der VEEN, G. (2017). Evaluation of the State Archives (ARA-AGR) of Belgium. Management summary. Rapport, Technopolis Group. [En ligne], https://www.belspo.be/belspo/fsi/doc/Peer_Review_ARA_Management_Summary.pdf, consulté le 16.03.2018, 13 p.

Annexes

Annexe 1
Inventaire de la documentation Wikibase

Cette annexe propose un inventaire aussi exhaustif que possible de l'ensemble des sources en ligne documentant le fonctionnement du logiciel Wikibase en date du mois de juin 2020.

— Wikibase [79]
— LearningWikibase [80]
— Phabricator, Wikibase-Containers [81]
— Wikibase Community User Group [82]
— Telegram group [83]
— Wikibase Community User Group - *Mailing list* [84]
— MediaWiki - Wikibase FAQ [85]
— MediaWiki - Wikibase Installation [86]
— Contact the development team [87]
— Using OpenStack to run a custom Wikibase [88]
— Wikibase for Research Infrastructure — Part 1 [89]
— Wikibase Install Basic Tutorial [90]
— Retours d'exploration Wikibase 2019 [91]
— Running and querying my own Wikibase instance [92]
— Wikibase : configure, customize, and collaborate [93]
— Installing Blazegraph and Wikibase [94]
— 2 minutes on installing Wikibase [95]
— Meta-Wiki - Wikibase Upgrade Workflow [96]
— Wikidata - Wikibase documentation [97]

79. https://wikiba.se/
80. http://learningwikibase.com
81. https://phabricator.wikimedia.org/project/profile/3079/
82. https://meta.wikimedia.org/wiki/Wikibase_Community_User_Group
83. https://t.me/joinchat/HGjGexZ9NE7BwpXzMsoDLA
84. https://lists.wikimedia.org/mailman/listinfo/wikibaseug
85. https://www.mediawiki.org/wiki/Wikibase/FAQ
86. https://www.mediawiki.org/wiki/Wikibase/Installation
87. https://www.wikidata.org/wiki/Wikidata:Contact_the_development_team
88. https://fuga.cloud/labs/using-openstack-to-run-custom-wikibase/
89. https://link.medium.com/wH6OkVlvJ6
90. https://semlab.io/howto/wikibase_basic
91. https://www.mediawiki.org/wiki/User:Jumtist/exploration_wikibase#Retours_d'exploration_Wikibase_2019
92. http://www.snee.com/bobdc.blog/2018/06/running-and-querying-my-own-wi.html
93. https://stuff.coffeecode.net/2018/wikibase-workshop-swib18.html
94. https://heardlibrary.github.io/digital-scholarship/lod/install/
95. https://youtu.be/P174BEDhUJg
96. https://meta.wikimedia.org/wiki/File:Wikibase_Upgrade_Workflow.pdf
97. https://www.wikidata.org/wiki/Wikidata:Wikibase_documentation

— GitHub repository : wikibase-docker [98]
— Addshore's blog *(posts published with a* Wikibase *tag)* [99]

98. https://github.com/wmde/wikibase-docker/blob/master/README.md
99. https://addshore.com/tag/wikibase/

Annexe 2
Analyse des requêtes d'utilisateurs dans le catalogue en ligne Pallas

Cette annexe présente les détails de l'analyse des requêtes d'utilisateurs effectuées dans Pallas, le catalogue en ligne du CegeSoma, au cours d'une période d'un an (entre le 1er septembre 2015 et le 31 août 2016).

Collecte des données

La première étape concerne la **collecte des données d'analyse**. Les données d'usage de Pallas ont été collectées par le centre de recherche iMec [100] entre septembre 2015 et avril 2017, à l'aide du logiciel libre Piwik [101]. Concrètement, il s'agit d'insérer un tag Javascript sur chaque page web, qui sera activé à chaque visite effectuée sur le catalogue Pallas et permettra au serveur Piwik d'enregistrer cette visite et de stocker les données correspondantes dans une base de données dédiée. L'identification des utilisateurs, qui est limitée à un identifiant numérique anonymisé, est basée soit sur des cookies stockés sur leur ordinateur [102], soit sur leur adresse IP. Bien qu'il soit également possible d'utiliser les logs du système de gestion de base de données pour obtenir les requêtes d'utilisateurs, il est apparu que ces derniers n'avaient pas été stockés de façon persistante par le CegeSoma. De plus, les données conservées par Piwik ne nécessitent pas d'étapes préliminaires visant à recréer les sessions de visite de chaque visiteur, contrairement aux logs (Nouvellet *et al.*, 2017). Enfin, Piwik, en combinant adresses IP et cookies HTTP, fournit un nombre de visiteurs plus précis que ne le permettent les fichiers log stockant uniquement des adresses IP [103].

Outre la possibilité de rester propriétaires des données brutes, Piwik propose une fonctionnalité pour suivre des actions spécifiques sur un site web

100. L'un des partenaires du projet MADDLAIN, anciennement connu sous le nom de iMinds
101. Connu sous le nom de Matomo depuis début 2018, Piwik représente une alternative *open source* à Google Analytics. La principale différence entre ces deux fournisseurs de service est que Piwik permet aux utilisateurs de stocker les données brutes, alors que Google reste propriétaire de ces données.
102. Les cookies sont des petits fichiers textuels qui sont placés sur le disque dur des utilisateurs, ils permettent d'identifier une connexion entre un navigateur web et un serveur web afin d'assurer ou d'améliorer la visualisation de pages web.
103. Une analyse similaire que nous avons effectuée dans le contexte d'une autre publication (Chardonnens *et al.*, 2018) a montré que près de 18% de visiteurs distincts identifiés par Piwik n'avaient pas été reconnus comme tels par les *log files*, le nombre de visiteurs pour une même période de test s'élevant respectivement à 1 071 pour les *log files* (identification basée sur des adresses IP) et 1 298 pour les données brutes Piwik (identification basée sur l'utilisation conjointe d'adresses IP et de cookies). En effet, comme l'a montré Voorbij (2010) : *IP addresses are limited when people use a collective access through a proxy server or when they operate in an environment that makes use of dynamic IP addresses.*

Visiteur ID	03A00AE7786DDA4E
Visite ID	735415
Heure	2016-01-12 04 :03 :15
URL	pallas.cegesoma.be/pls/opac/plsp.getplsdoc?lan=F&htdoc=general/opac.htm/
Variable personnalisée	FormDatatext=RADIOBRUXELLES&action=search&Seop=6&in=_BA

TABLE 2 – Exemple de données collectées par le logiciel Piwik (les termes de requête ont été mis en évidence à l'aide de majuscules).

à l'aide de variables personnalisables. Ces variables sont très utiles lorsque les URLs ne sont pas explicites. En effet, l'une des particularités du catalogue Pallas est qu'il affiche des URLs opaques. Cela signifie que lorsqu'une recherche est effectuée, l'URL reste toujours la même : http://pallas.cegesoma.be/pls/opac/plsp.getplsdoc?lan=F&htdoc=general/opac.htm. Il n'y donc aucune trace dans l'URL des termes recherchés. Pour pouvoir explorer la façon dont les utilisateurs recherchent des contenus, il a fallu ajouter des paramètres à ces URLs en les personnalisant. Elles ont été « augmentées » à l'aide de variables basées sur du code Javascript et enregistrées automatiquement par Piwik [104]. Le résultat stocké dans les données brutes ne présente donc pas de résultats directement exploitables [105]. Il faut utiliser des méthodes computationnelles pour pouvoir récupérer les informations pertinentes. La première étape consiste à formuler une requête SQL permettant d'extraire de la base de données Piwik les données brutes correspondant aux critères souhaités. Concrètement, le jeu de données utilisé dans le cadre de cette analyse couvre une période d'une année, allant du 1er septembre 2015 au 31 août 2016. Outre l'identifiant du visiteur et de la visite, les autres éléments extraits sont l'horodatage, l'URL, ainsi que la variable personnalisée contenant – notamment – les termes de recherche de l'utilisateur. Le tableau 2 illustre les cinq types d'éléments extraits de cette base de données.

104. Ce processus a été réalisé manuellement, au cas par cas, voir Chardonnens *et al.* (2017).

105. Il faut toutefois souligner que Piwik offre des fonctionnalités permettant d'enregistrer et stocker des mots-clés issus de recherches internes et d'obtenir ces derniers plus facilement. Bien que ces fonctionnalités n'aient pas été exploitées dans le cadre du projet MADDLAIN, elles pourraient potentiellement faciliter les étapes de pré-traitement détaillées dans les paragraphes suivants.

Pré-traitement des données brutes

La seconde étape concerne les diverses tâches de pré-traitement des données brutes. En effet, comme illustré à l'aide du tableau 2, les termes de recherche sont stockés à l'aide d'une variable personnalisée. Cette variable peut comporter plusieurs valeurs ; ainsi, outre les termes de recherche, elle contient également d'autres informations sur les choix du visiteur au cours de sa recherche, comme par exemple le fait qu'il souhaite utiliser un filtre et effectuer une recherche exclusivement dans les archives ou photographies. Cela signifie qu'il est nécessaire de distinguer les éléments pertinents pour les isoler du reste. Par ailleurs, il arrive que les termes de la requête soient rémanents et continuent à figurer dans la variable personnalisée sans qu'une nouvelle recherche n'ait pour autant été effectuée, par exemple si une utilisatrice décide de consulter l'arbre d'archives proposé par Pallas [106]. Par ailleurs, il ne semble pas non plus pertinent de compter à plusieurs reprises les termes *école Mons* et *ecole mons* si seules de légères différences formelles (l'accent aigu et la majuscule) les distinguent. L'étape de pré-traitement vise donc à gérer ce type de tâches afin d'avoir des données prêtes à être analysées.

Concrètement, ces tâches sont au nombre de quatre et ont été effectuées à l'aide d'une bibliothèque Python d'analyse de données : Pandas. Elles permettent à la fois de limiter le nombre de données à analyser et de travailler avec des données aussi significatives que possible. La première tâche, le *parsing*, vise à extraire dans une colonne dédiée les termes de recherche en faisant appel à des expressions régulières. La seconde tâche a pour but de traiter les termes de recherches entrés une seule fois au cours d'une visite mais apparaissant à plusieurs reprises. Sachant que prendre en compte ces occurences multiples pourraient biaiser des résultats comme les termes les plus fréquents ou le nombre de noms de personnes, une méthode approximative et perfectible, mais néanmoins consistante et homogène, est adoptée. Le principe, implémenté à l'aide de l'opération Pandas *group by*, est de ne garder chaque requête distincte qu'une seule fois par visite. La troisième tâche consiste à « nettoyer » les requêtes des utilisateurs de manière à pouvoir associer des chaînes de caractères similaires en dépit de différences superficielles (Schütze *et al.*, 2008). Pour détecter ces « doublons cachés », différentes mesures peuvent être prises : supprimer des espaces de début et de fin ou des espaces consécutives, convertir tous les caractères en bas de casse, remplacer tout caractère qui n'est pas alphanumérique par une espace, ou encore remplacer les caractères spéciaux (par exemple, $\tilde{A}$ devient un simple a). En outre, les requêtes contenant uniquement des chiffres, de

106. Dans la variable *FormData*, nous retrouverons ainsi successivement *radiobruxelles&action=SEARCH&Seop=6&in=_BA* et *radiobruxelles&action=BROWSE-ARCHIVES&Seop=6&in=_BA*.

même que celles contenant moins de trois caractères ont été jugées peu significatives et dès lors ôtées du corpus. Enfin, la dernière tâche consiste à reproduire l'opération de regroupement des requêtes identiques, en partant cette fois-ci des termes normalisés [107].

À la fin de cette étape de pré-traitement, le jeu de données contient un total de 30 703 requêtes, parmi lesquelles 21 385 requêtes distinctes ont été identifiées. 9 253 visites sont à l'origine de ces 30 703 requêtes. Le minimum est de 1 requête par visite et le maximum est de 132 requêtes par visite. Chaque requête consiste en un ou plusieurs *tokens* [108].

Extraction des noms de personnes

La troisième étape a pour objet l'extraction de potentiels noms de personnes. Si l'extraction d'entités nommées a fait l'objet de nombreuses recherches et que divers outils ont fleuri sur le web afin de faciliter cette opération, il faut garder à l'esprit que notre corpus a des propriétés particulières. En effet, il ne s'agit non pas de textes rédigés, dotés de majuscules, structures grammaticales et signes de ponctuations, mais d'un corpus composé de textes non structurés, ambigus, très courts. Les services web « généralistes » de *Named Entity Recognition* ne semblent donc pas adaptés à notre corpus.

Par chance, une analyse portant sur un corpus similaire [109] a été réalisée collaborativement avec d'autres chercheurs du centre de recherche ReSIC [110] et a permis, d'une part, de tester les taux de succès de ces services web, et, d'autre part, de développer en interne un outil plus adapté à la nature particulière du corpus. En ce qui concerne les services web, des tests préliminaires ont été effectués sur sept d'entre eux [111] à l'aide d'un *Gold Standard Corpus*. Ce dernier est basé sur un échantillon de 1 000 requêtes annotées manuellement par deux chercheurs. Au terme d'une recherche de consensus entre les deux annotateurs, 473 requêtes ont été annotées comme *Personne*. Sur ces 473 entités, seules 141 contenaient une structure correspondant à un « nom complet », c'est-à-dire un (supposé) prénom accompagné d'un (supposé) nom de famille. Le service Web qui proposait les résultats les plus convaincants, Rosette, a identifié correctement 128 noms de personnes sur

107. Cette méthode de pré-traitement des données a été décrite de façon exhaustive dans une étude de cas similaire, voir Chardonnens et Hengchen (2017)

108. Nous définissons *token* comme une chaîne de caractères *not containing any non-printable or delimiting characters (blank, tabulator, line-feed, new line, etc.* (Hassler et Fliedl, 2006).

109. Les requêtes d'utilisateurs de l'application web BelgicaPress – mise en ligne par KBR, la bibliothèque nationale de Belgique.

110. Nous remercions Seth van Hooland et plus particulièrement Ettore Rizza, qui a pris en charge le volet technique.

111. Rosette, Dandelion, Babelfy, TagMe, Dexter, DBpedia Spotlight et le Stanford NE Recogniser entraîné sur l'anglais

ces 141, soit 90,8%. Afin d'obtenir un meilleur taux de rappel, il a été jugé plus stratégique de développer en interne un outil tenant compte de la nature particulière des requêtes d'utilisateurs.

Cet outil est composé d'un script Python basé sur des listes et des règles linguistiques devant permettre d'extraire de façon automatisée des noms de personnes [112]. Partant du principe qu'un prénom ou un nom de famille seuls ne suffisent pas à identifier une personne [113], nous avons choisi d'extraire uniquement les (potentiels) noms de personnes « complets » (c'est-à-dire composés d'un prénom ET d'un nom de famille). L'objectif est de privilégier le rappel plutôt que la précision, étant donné que les résultats obtenus sont destinés à être affinés au cours d'une étape ultérieure faisant appel à des bases de connaissance et permettant d'isoler les (potentiels) faux positifs.

Voici les principales étapes qui ont été « traduites » en fonctions Python :

1. Identifier les requêtes composées de plus d'un *token*.
2. Repérer parmi tous ces noms potentiels ceux qui contiennent un prénom [114].
3. Une fois le (supposé) prénom identifié, déduire à l'aide de quelques règles linguistiques quelle partie de la requête constitue probablement le nom de famille.

Cet extracteur a été amélioré au cours d'un processus itératif, permettant d'adapter les listes et règles aux données concernées. Ainsi, il a par exemple fallu adapter les règles linguistiques pour tenir compte des noms de famille incluant une particule comme van, von ou van den. Par ailleurs, la liste de prénoms a dû être adaptée au contexte du CegeSoma lorsque le script a été lancé sur les requêtes Pallas : en effet, s'il arrive que France corresponde parfois à un prénom, il est apparu qu'il entraînait trop de faux positifs dans le contexte du CegeSoma. Si d'autres prénoms « problématiques » ont dû être enlevés de la liste de références, d'autres mots ont dû quant à eux être ajoutés, comme par exemple « Général », contenu dans près d'une centaine de requêtes, et participant à désigner un individu, comme par exemple *Général Armengaud*. Ces ajouts et suppressions ne vont pas sans la perte de vrais positifs ou l'ajout de faux positifs. C'est par exemple le cas de requêtes contenant les termes « Général lieutenant » ou *Général aviation*, qui ne désignent pas une personne en particulier, mais seront considérés comme tels par l'extrac-

112. Notons qu'il a également été paramétré pour pouvoir extraire des noms de lieux en Belgique, pour des détails, voir Chardonnens *et al.* (2018).

113. À part dans le cas où ce sont des personnalités très connues comme par exemple l'auteur belge de bandes dessinées Hergé.

114. Cette étape a été réalisée à l'aide d'une liste de référence de plus de 60 000 prénoms, constituée à partir de plusieurs jeux de données disponibles en ligne ou au sein de l'institution.

teur. Malheureusement de tels cas sont inévitables sans un contrôle manuel systématique [115].

Lancé sur les 30 703 requêtes « normalisées » – correspondant à 21 385 requêtes distinctes –, ce script d'extraction a permis d'identifier 4 104 requêtes contenant possiblement des noms de personnes « complets », correspondant à 2 885 chaînes de caractères distinctes.

Réconciliation des noms

Enfin, après la collecte des données, leur pré-traitement et l'extraction des noms de personnes, la dernière étape a pour objectif la réconciliation des (supposés) noms de personnes avec des bases de connaissance susceptibles de les identifier. Cette opération, appelée *Named Entity Linking* (Rao *et al.*, 2013), vise à associer chaque entité nommée apparaissant dans une requête à son entité correspondante dans une base de connaissance. Deux bases de connaissance ont été utilisés dans le cadre de ce travail : une généraliste, Wikidata, et une plus spécialisée, Viaf [116]. Précisons qu'un test préliminaire a également été effectué avec le services de réconciliation ULAN (*Union List of Artist Names*) du Getty Research Institute, ainsi qu'avec le service de réconciliation SNAC (*Social Netowrk Archival Context*). Les efforts n'ont cependant pas été approfondis en raison de l'important bruit contenu dans les résultats [117] ainsi que du manque de pertinence dans l'attribution des *perfect match* [118].

Cette utilisation conjointe de VIAF et Wikidata vise à augmenter le nombre de noms réconciliés. En effet, une comparaison avec les statistiques de Wikidata révèle que VIAF contient près de deux fois plus de noms de personnes. En revanche, le périmètre couvert par ces deux bases de connaissance est différent, VIAF étant associé au domaine du patrimoine culturel, tandis que Wikidata a une vocation beaucoup plus généraliste (*the sum of all human knoweldge*). Cela semble intéressant dans le contexte qui nous intéresse, étant

115. Il s'agit donc d'effectuer un arbitrage en évaluant les coûts et bénéfices de chaque approche, pour déterminer au cas par cas ce qui semble le plus efficace et le moins dommageable. Ce sont des éléments à garder en tête au moment d'énoncer des résultats et de formuler des conclusions : ce processus d'extraction semi-automatisée permet certes de traiter rapidement plusieurs milliers de données, en revanche, il nécessite de faire le deuil de la précision et de l'exactitude, pour privilégier des affirmations reflétant une réalité nuancée et approximative.

116. Pour des détails sur leurs caractéristiques et différences, veuillez consulter (Chardonnens *et al.*, 2018).

117. Ainsi, le service ULAN propose par exemple pas moins de 25 candidats pour les termes *Karl Marx*, incluant des noms éloignés tels que *Burle Marx, Roberto*, mais également des institutions telles que *Karl-Marx-Universität, Archäologiesches Institut*.

118. Ainsi dans le cas du service SNAC, des requêtes comme *Albert Seghers* ; *Amandine Dumon* ou *Alois Pollet* sont erronément associées à un *perfect match*, à savoir une entité SNAC contenant uniquement un nom de famille (*Seghers* ; *Dumon* ; *Pollet*).

donné que les collections du CegeSoma couvrent des questions de société qui dépassent largement le cadre du patrimoine culturel. La confrontation des résultats obtenus à l'aide de ces deux sources devrait permettre de visualiser leur complémentarité.

Pour mener à bien cette dernière étape de réconciliation, les potentiels noms de personnes ont été traités à l'aide du logiciel OpenRefine et des services web de réconciliation basés sur les APIs de Viaf et de Wikidata [119]. OpenRefine propose deux fonctionnalités intéressantes : d'une part, le logiciel permet de sélectionner un type auquel associer le nom à réconcilier (dans notre cas, nous recherchons des êtres humains) afin d'accélérer le processus et de diminuer l'ambiguïté ; d'autre part, à chaque nom *reconnu* par Viaf ou Wikidata est associé une liste de *candidats à la réconciliation*, dotés d'un score indiquant le taux de similarité. Ainsi, si une chaîne de caractères telle que *andre cauvin* – faisant référence au réalisateur belge André Cauvin – peut « facilement » être associée à l'entité Wikidata correspondante Q2847459|André Cauvin [120], d'autres chaînes de caractères plus ambigües, comme *Hermann Schmidt*, ne permettent pas une réconciliation par défaut. En effet, *Hermann Schmidt* s'avère être un nom porté par une bonne vingtaine d'entités Wikidata : sans information contextuelle, impossible de savoir à quel individu la requête faisait référence. Dans ce genre de cas, où plusieurs candidats possèdent un score de 100%, nous nous contentons d'en déduire qu'il est plus que probable qu'il s'agisse d'un véritable nom de personne, sans chercher à réconcilier ce nom avec l'entité la plus plausible [121].

Dans d'autres cas, aucune chaîne de caractères ne correspond strictement à ce qui a été encodé dans la requête : une liste de noms candidats est alors proposée avec des scores de similarité moins élevés, laissant à l'utilisateur le soin d'effectuer les vérifications manuelles pour savoir si *Henri Man* correspond éventuellement à Q666890|Henri De Man [122], qui est par ailleurs également connu comme Hendrik de Man. Dans le cadre d'un alignement entre divers fichiers d'autorité, destiné de surcroît à être réutilisé par d'autres utilisateurs, de telles vérifications seraient bien entendu cruciales et devraient être complétées à l'aide de données additionnelles permettant de lever l'ambiguïté. En revanche, dans notre cas visant avant tout à identifier quelle proportion des requêtes effectuées dans un catalogue en ligne

119. En présence d'un jeu de données plus important susceptible de mettre à mal les capacités d'OpenRefine, cette opération aurait également pu être réalisée à l'aide d'un langage de programmation tel que Python, voir par exemple ce script : https://github.com/mnyrop/pywyky.

120. https://www.wikidata.org/wiki/Q2847459.

121. Cette étape s'avérerait toutefois utile si nous souhaitions poursuivre l'analyse en déduisant de nouvelles informations à l'aide des entités liées, comme par exemple le pays de nationalité des personnes recherchées, leur occupation ou encore leur genre.

122. https://www.wikidata.org/wiki/Q666890.

contiennent des noms de personnes, cette exactitude n'est pas requise et serait même contre-productive : tout le temps gagné à l'aide de processus semi-automatisés serait perdu en devant effectuer plusieurs milliers de vérifications manuelles. Nous avons donc opté pour un compromis : parmi tous les potentiels noms de personnes, seuls les noms ayant obtenu un score de réconciliation au-dessus d'un certain seuil sont retenus.

Avant de présenter les résultats obtenus au terme du processus de réconciliation, nous aimerions souligner que la façon d'attribuer des scores varie en fonction du service de réconciliation, comme le montre la figure 1. Ainsi, si l'on prend l'exemple de Karl Marx, Wikidata nous propose sept *perfect match* à 100%, correspondant à diverses entités portant le nom de *Karl Marx* (ainsi que trois autres candidats avec des scores inférieurs, allant de 62 à 90%), tandis que Viaf nous propose un *perfect match* à 100% [123], tandis que deux autres candidats avec des scores largement inférieurs sont proposés : Lenin, Vladimir Il'ich 1870-1924 à 21%, ainsi que Stalin, Joseph 1878-1953 à 8%) [124]. Cela signifie, d'une part, que cela aurait peu de sens de fixer un même seuil de score à atteindre pour chacun des services de réconciliation et, d'autre part, qu'il faut garder à l'esprit que plus une ressource externe est riche, plus elle est susceptible de contenir des homonymes (comme par exemple plusieurs Karl Marx), rendant ainsi les déductions moins évidentes.

D'autre part, se pose la question de la façon dont l'ordre des éléments et la structure du nom impactent les scores de réconciliation. En effet, Viaf propose par exemple des noms répondant à la structure *prénom nom* sans virgule, comme par exemple *Léon Sarteel*. Cet exemple semble judicieux étant donné que ce sculpteur belge est présent à la fois dans Wikidata et VIAF, sans pour autant posséder d'homonymes connus, ce qui permet d'observer (voir tableau 3) comment le type d'algorithme utilisé par le service de réconciliation conditionne le score obtenu, en fonction de l'ordre des éléments et enfin de la présence d'une virgule ou d'un accent sur le *e* de Léon dans la chaîne de caractères soumise au processus de réconciliation. Étant donné l'impact manifeste de l'ordre des éléments sur les résultats de la réconciliation pour VIAF, nous avons dès lors décidé de procéder à une seconde vague de réconciliation pour ce service, après avoir isolé les *perfect matches* obtenus lors de la première étape et surtout inversé l'ordre des prénoms et noms.

123. Techniquement, le service de réconciliation utilisé pour Viaf donne des scores sur 1, convertis ici en pourcentages dans un souci d'homogénéisation.

124. Il est par ailleurs assez curieux et intriguant de relever qu'une recherche effectuée directement sur le site web de Viaf (en entrant *Karl Marx* comme termes de recherche, avec le filtre *noms de personne*, et *tout VIAF* comme fichier source, 1 139 résultats sont proposés, Vladimir Il'ich Lenin arrivant en première position, Joseph Stalin en seconde, et Karl Marx à la troisième place seulement. Voilà qui illustre parfaitement l'importance d'avoir des processus de recherche transparents et cohérents plutôt que des *black box* laissant l'utilisateur ignorant des processus utilisés.

FIGURE 1 – Exemple de scores de réconciliation Wikidata, Viaf et Ulan pour *Karl Marx*. Source : capture d'écran, logiciel OpenRefine.

Nom à tester	**Score Wikidata**	**Score Viaf**
Leon Sarteel	100	92
Leon, Sarteel	100	85
Léon Sarteel	100	100
Léon, Sarteel	100	92
Sarteel Leon	100	17
Sarteel, Leon	100	15
Sarteel Léon	100	17
Sarteel, Léon	100	15

TABLE 3 – Test d'impact de la modification de la graphie d'un nom sur les scores de réconciliation (pour des raisons d'harmonisation, les scores ont été convertis en pourcentages et arrondis à l'entier supérieur lorsque c'était nécessaire).

Enfin, voici les résultats obtenus au cours de l'étape de réconciliation de ces 2 865 chaînes de caractères (constituant de probables noms de personnes) :

— Avec Wikidata : 648 correspondances considérées comme parfaites (score entre 96[125] et 100%), et 657 autres candidats obtenus via l'utilisation conjointe de deux filtres (premièrement, l'entité Wikidata candidate doit obligatoirement être une personne, et deuxièmenent, le score doit être supérieur ou égal à un seuil minimal de 50%[126]), soit un total de 1 293 candidats.

— Avec VIAF : aux 193 correspondances parfaites (score de 1, équivalent à 100%) issues de la *première vague* de réconciliation, s'ajoutent, après l'inversion de l'ordre des prénoms et noms – pour mieux correspondre à la structure de données de VIAF, comme précisé au cours du paragraphe précédent –, 4 nouvelles correspondances parfaites, ainsi que 811 candidats dont le score est supérieur ou égal à un seuil de 33% (en réalité 0,33)[127], soit un total de 1 008 candidats.

Enfin, comme le montre la figure 2, plus de la moitié (53,41%) de ces 2 885 potentiels noms de personnes a pu être réconciliée avec Wikidata et, ou Viaf, dont un quart (25,65%) a été associé à la fois à une entité Wikidata et une entité VIAF. Le quart restant (27,77%), composé de requêtes ayant été associées à l'une des ressources seulement, confirme la complémentarité de ces deux bases de connaissance dans le cadre de cette analyse.

125. Par exemple Hannah Arend - Hannah Arendt.

126. Un survol des candidats avec des scores inférieurs à 50% a montré qu'il s'agissait principalement de faux positifs (comme par exemple Nice France), tandis que les scores supérieurs à 50% s'apparentent davantage à des noms de personne, bien que l'entité Wikidata ne corresponde pas parfaitement).

127. Ce seuil a été fixé manuellement après survol des données. S'il peut sembler peu élevé de prime abord, il s'est avéré qu'il permet d'intégrer des correspondances correctes mais bénéficiant d'un score peu élevé, comme par exemple la requête *Jongh Andree* qui est assimilé à 42% seulement avec l'entité VIAF *De Jongh, Andrée 1916-2007*.

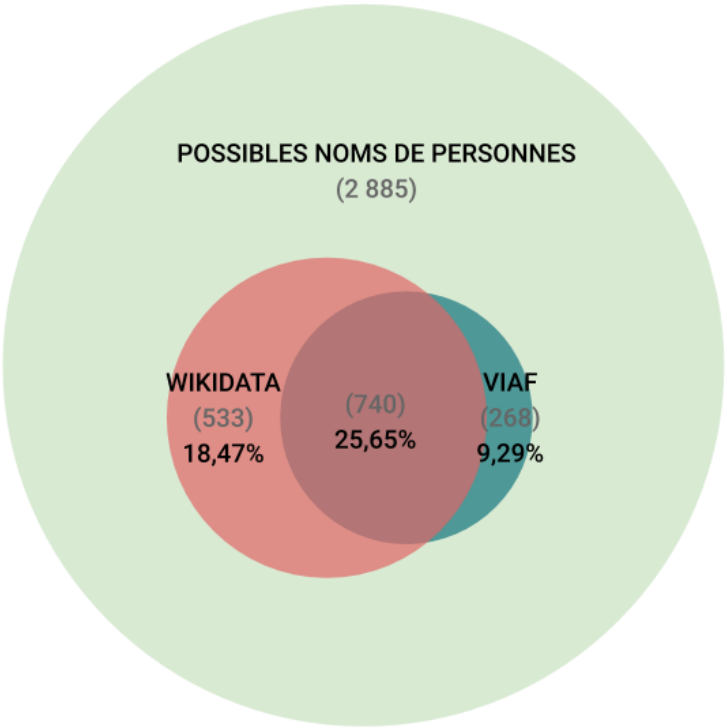


FIGURE 2 – Aperçu de la répartition des réconciliations des (possibles) noms de personnes avec Viaf et Wikidata (diagramme à l'échelle).

Annexe 3
Alignement des noms de lieux situés en Belgique avec Wikidata

Cette annexe présente en détail les efforts déployés pour enrichir le référentiel des Archives de l'État portant sur les lieux situés en Belgique. Ces efforts ont résulté en un enrichissement mutuel de Wikidata et du référentiel.

Le fichier de départ, mis à notre disposition par le gestionnaire des bases de données des Archives de l'État, contient plus de 3 500 noms de lieux (communes, sections de communes, hameaux et localités) situés en Belgique, accompagnés de leurs coordonnées géographiques ainsi que de leur code INS [128].

Notre objectif était d'aligner ces entités géographiques avec des éléments Wikidata. Cette étape a impliqué d'effectuer une réconciliation en utilisant les codes INS comme pivots entre les deux jeux de données. En effet, comme illustré au travers de notre thèse, un identifiant numérique ne présente pas les mêmes difficultés que des chaînes de caractères, sujettes à l'ambiguïté. La difficulté est venue du fait que le code INS n'était pas documenté sur la plupart des fiches Wikidata des communes ou sections de communes belges. Une phase préliminaire a ainsi consisté à compléter les fiches Wikidata pour y faire figurer leur code INS.

Cette démarche annexe de complétion des fiches Wikidata a été réalisée en commençant par extraire – à l'aide d'une requête SPARQL [129] — toutes les entités Wikidata ayant comme P31|nature : Q493522|commmunes ou Q2785216|sections de commune belges, avant de les réconcilier avec les données mises à disposition par Statbel [130]. Ce processus, réalisé de façon semi-automatisée à l'aide du service de réconciliation OpenRefine « Reconcile CSV » [131] a mis en lumière sept types de difficultés :

— la coexistence d'anciens et de nouveaux codes INS [132]
— les récentes fusions de communes n'ayant pas encore été prises en compte sur Wikidata [133]
— la présence d'homonymes [134] nécessitant un travail de désambiguïsation supplémentaire

128. Ce code numérique composé de cinq caractères est attribué à chaque entité administrative belge par Statbel, l'office belge de statistique (Statbel, 2018).
129. https://w.wiki/6rn.
130. Voir Statbel (2019).
131. http://okfnlabs.org/reconcile-csv/.
132. C'est le cas par exemple des arrondissements administratifs de la province du Hainaut, voir : Statbel (2018).
133. Sans parler de nouveaux noms comme Kruisem, issus de la fusion de Kruishoutem et Zingem.
134. Par exemple St-Nicolas, Alost ou Halle.

— la présence de faux positifs, due à l'usage d'un algorithme de *fuzzy matching* [135]
— les cas délicats de codes INS englobant plusieurs noms de lieux [136]
— le manque de cohérence et, ou de complétude sur Wikidata [137]
— le niveau de granularité très fin de Statbel qui ne correspond pas toujours à celui de Wikidata [138].

Cette opération a donc nécessité plusieurs étapes successives de contrôle et de vérification, afin d'identifier les éventuelles erreurs et d'améliorer le nombre d'éléments réconciliés en tenant compte des constats listés ci-dessus. Il s'agissait notamment de faire appel à la logique hiérarchique, voulant que chaque commune ou section de commune puisse être associée à une entité administrative supérieure (commune ou province). Comme par exemple Q2687317|Chevetogne [139], qui fait P361|partie de Q456490|Ciney [140]. Mais là encore, il a fallu tenir compte du contexte belge, du multilinguisme et des fusions de communes : le fichier de Statbel n'est pas multilingue, or, les chaînes de caractères *Doornik* et *Tournai* ne semblent rien avoir en commun (alors qu'il s'agit respectivement des formes en néerlandais et en français d'une même commune belge) ; Hansbeke, ancienne section de Nevele, fait désormais partie de Deinze, étant donné que Nevele a fusionné le 1er janvier avec Deinze et changé de nom, mais l'information n'est pas encore à jour sur Wikidata, etc. Au final, cette étape a conduit à la création de fiches Wikidata pour 280 sections de communes manquantes et à l'ajout de plus de 2 100 codes INS [141] pour des communes et sections de communes belges.

Cette étape terminée, il a été possible d'utiliser le code INS comme pivot pour réconcilier les données du référentiel des Archives de l'État avec les données de Wikidata. Cela n'a pas été possible dans l'intégralité des cas, certaines occurences particulières nécessitant une exploration plus poussée ou des traitements sur mesure, mais cela a néanmoins permis de réconcilier le référentiel des Archives avec 2 812 identifiants Wikidata. Une fois le

135. Comme Beerse et Beersel.
136. C'est le cas par exemple d'anciennes communes qui ont été divisées en *parties de* … Ainsi une commune comme Oupeye compte, entre autre, une « partie de Hersal » alors que Wikidata possède une contrainte spéficiant qu'un seul code INS ne devrait être utilisé par entité Wikidata, et que chaque code INS ne devrait être utilisé qu'une seule fois.
137. Certaines entités Wikidata semblent correspondre à des sections de commune mais ne sont pas strictement désignées comme *section de commune*.
138. Les attributions de codes INS de Statbel se calquant sur des réalités administratives, elles ne sont pas toujours en adéquation avec les pratiques en usage sur Wikidata, c'est le cas par exemple pour Anvers et ses multiples *Administrat. wijk of distr.* ne possédant pas de nom distinct et qui véhiculeraient du bruit dans les recherches Wikidata si une fiche était créée pour chacune d'entre elles.
139. https://www.wikidata.org/wiki/Q2687317.
140. https://www.wikidata.org/wiki/Q456490.
141. À l'aide de la propriété Wikidata P1567|code INS.

lien créé, il a été aisé d'extraire d'autres informations de Wikidata, comme par exemple 1 032 identifiants GeoNames, ou encore les libellés Wikidata en français et en néerlandais, afin d'enrichir le référentiel des Archives. Cependant, après exploration approfondie, il s'est avéré que ce dernier n'est toutefois pas exempt de quelques fantaisies : il s'avère en effet qu'il contient quelques « pseudo communes [créées] lorqu'une paroisse était sur plus d'une commune (cela fait partie des bricolages) »[142]. Sachant que laisser ces entités signifierait rajouter du bruit et prendre le risque de flouter lors des résultats de processus de réconciliation, nous avons pris le parti de les retirer, lorsqu'elles ont pu être identifiées[143].

142. Échange de mail avec Yves Lardinois, 13.01.2020.
143. Ce qui n'est pas chose aisée, étant donné que rien ne distingue formellement ces *inventions* de véritables communes fusionnées comme Céroux-Mousty.

Annexe 4
Entity linking

Record linkage

In [2]:

```
import pandas as pd
import sys
import recordlinkage
from recordlinkage.preprocessing import clean
from recordlinkage.preprocessing import phonetic
sys.executable
```

Out[2]:

```
'/Library/Frameworks/Python.framework/Versions/3.7/bin/python3'
```

0. Preprocessing

In [5]:

```
df11 = pd.read_csv('data/20200210_belgiumV4.csv', sep=',', header='infer')
df11.head(2)
```

Out[5]:

	B_name	B_prenom	B_nom	B_trfwnumm	B_birth_year	B_death_year	B_wikidata_id
0	Achille Van Acker	Achille	Van Acker	NaN	1898.0	1975.0	Q1499
1	Adrien Emile Van Coppenolle	Adrien Emile	Van Coppenolle	NaN	1893.0	1975.0	NaN

In [6]:

```
df11 = df11.drop(columns=['Column'])
df11 = df11[['B_name','B_prenom','B_nom','B_birth_year','B_death_year', 'B_wikidata_id'
]]
```

In [7]:

```
df22 = pd.read_csv('data/20200218_adieuxV1.csv', sep=',', header='infer')
#df22[['P_trfwnumm']] = df22[['P_trfwnumm']].astype(float)
```

In [9]:

```
df22.head(2)
```

Out[9]:

	A_name	A_nom	A_prenom	secondnames	cleaned_place	birth	death	Notes
0	Jean Ackermans	ackermans	jean	NaN	NaN	NaN	1943.0	NaN
1	Henri Aerden	aerden	henri	NaN	anvers	1897.0	1942.0	NaN

In [10]:

```
df11["cleaned_prenom"]= clean(df11["B_prenom"],lowercase=True, replace_by_whitespace='[\\-\\_]', strip_accents='unicode', remove_brackets=True, encoding='utf-8', decode_error='ignore')
df11["cleaned_nom"]= clean(df11["B_nom"],lowercase=True, replace_by_whitespace='[\\-\\_]', strip_accents='unicode', remove_brackets=True, encoding='utf-8', decode_error='ignore')
df22["cleaned_prenom"]= clean(df22["A_prenom"],lowercase=True, replace_by_whitespace='[\\-\\_]', strip_accents='unicode', remove_brackets=True, encoding='utf-8', decode_error='ignore')
df22["cleaned_nom"]= clean(df22["A_nom"],lowercase=True, replace_by_whitespace='[\\-\\_]', strip_accents='unicode', remove_brackets=True, encoding='utf-8', decode_error='ignore')
df11.head(2)
```

Out[10]:

	B_name	B_prenom	B_nom	B_birth_year	B_death_year	B_wikidata_id	cleaned_pre
0	Achille Van Acker	Achille	Van Acker	1898.0	1975.0	Q14997	a
1	Adrien Emile Van Coppenolle	Adrien Emile	Van Coppenolle	1893.0	1975.0	NaN	adrien e

In [11]:

```
df11["phoneticPrenom"] = phonetic(df11["cleaned_prenom"], method="nysiis") # cf algorithmes ici : https://recordlinkage.readthedocs.io/en/latest/ref-preprocessing.html
df11["phoneticNom"] = phonetic(df11["cleaned_nom"], method="nysiis") # cf algorithmes ici : https://recordlinkage.readthedocs.io/en/latest/ref-preprocessing.html
df22["phoneticPrenom"] = phonetic(df22["cleaned_prenom"], method="nysiis")
df22["phoneticNom"] = phonetic(df22["cleaned_nom"], method="nysiis")
```

In [12]:

```
df11.columns = df11.columns.str.replace('B_birth_year','birth')
df11.columns = df11.columns.str.replace('B_death_year','death')

#df22.columns = df22.columns.str.replace('cleaned_birth_year','birth')
#df22.columns = df22.columns.str.replace('cleaned_death_year','death')
```

In [13]:

```
df11.head()
```

Out[13]:

	B_name	B_prenom	B_nom	birth	death	B_wikidata_id	cleaned_prenom	cl
0	Achille Van Acker	Achille	Van Acker	1898.0	1975.0	Q14997	achille	
1	Adrien Emile Van Coppenolle	Adrien Emile	Van Coppenolle	1893.0	1975.0	NaN	adrien emile	
2	Albert Lilar	Albert	Lilar	1900.0	1976.0	Q466832	albert	
3	Albert Servaes	Albert	Servaes	1883.0	1966.0	Q2197592	albert	
4	Alexander von Falkenhausen	Alexander	von Falkenhausen	1878.0	1966.0	Q62521	alexander	fa

1. indexing

In [14]:

```
from recordlinkage.index import Full
```

/!\ Blocking

In [29]:

```
# Create indexing object
indexer = recordlinkage.index.Block(left_on='cleaned_nom')
pairs_block2 = indexer.index(df11, df22)
#indexer = recordlinkage.index.Block(left_on='cleaned_name')

from recordlinkage.index import SortedNeighbourhood
indexer = recordlinkage.SortedNeighbourhoodIndex('cleaned_nom', window=9)
pairs_neighbour = indexer.index(df11, df22)
```

In [30]:

```
print (len(df11), len(df22), len(pairs_block2), len(pairs_neighbour))
```

```
88 381 5 599
```

2. COMPARE

In [31]:

```
#initialiser la classe
comp = recordlinkage.Compare()

comp.exact("cleaned_nom", "cleaned_nom", label="N_exact") #N= Nom
comp.string("cleaned_nom", "cleaned_nom", method='jarowinkler', threshold=0.60, label=
"N_jarowink")

comp.exact("cleaned_prenom", "cleaned_prenom", label="P_exact") #P= Prenom
comp.string("cleaned_prenom", "cleaned_prenom", method='jarowinkler', threshold=0.60, l
abel="P_jarowink")

comp.exact("phoneticPrenom", "phoneticPrenom", label="P_Phon_exact")
comp.exact("phoneticNom", "phoneticNom", label="N_Phon_exact")

#dates
#comp.date('birth64', 'birth64', label='date1A')
comp.exact('birth', 'birth', label='date1_exact')
comp.numeric('birth', 'birth', label='date1_numeric')
#comp.date('death64', 'death64', label='date2A')
comp.exact('death', 'death', label='date2_exact')
comp.numeric('death', 'death',label='date2_numeric')
```

Out[31]:

```
<Compare>
```

In [32]:

```
features2 = comp.compute(pairs_block2, df11, df22)
```

In [33]:

```
features3 = comp.compute(pairs_neighbour, df11, df22)
```

In [34]:

```
matches2 = features2[features2.sum(axis=1) > 2]
print(len(matches))
```

```
25
```

In [35]:

```
matches3 = features3[features3.sum(axis=1) > 2]
print(len(matches))
```

```
25
```

In [37]:

```
matches2['score'] = matches.apply(lambda x: x.sum(), axis=1)
matches3['score'] = matches.apply(lambda x: x.sum(), axis=1)
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packa
ges/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

In [38]:

```
matches3.head()
```

Out[38]:

		N_exact	N_jarowink	P_exact	P_jarowink	P_Phon_exact	N_Phon_exact	date1_exa
60	**28**	0	0.0	0	0.0	0	0	
66	**66**	0	1.0	0	0.0	0	0	
	65	0	1.0	0	0.0	0	0	
50	**170**	0	1.0	0	0.0	0	0	
61	**79**	0	1.0	0	0.0	0	0	

In [39]:

```
matches3.index.names = ['belgium', 'adieux']
matches3.to_csv('data/matches_lettresadieuxV1.csv')
```

In [40]:

```
matches3.head()
```

Out[40]:

		N_exact	N_jarowink	P_exact	P_jarowink	P_Phon_exact	N_Phon_exact	d
belgium	**adieux**							
60	**28**	0	0.0	0	0.0	0	0	
66	**66**	0	1.0	0	0.0	0	0	
	65	0	1.0	0	0.0	0	0	
50	**170**	0	1.0	0	0.0	0	0	
61	**79**	0	1.0	0	0.0	0	0	

In [43]:

```
df11.dtypes
```

Out[43]:

```
B_name             object
B_prenom           object
B_nom              object
birth             float64
death             float64
B_wikidata_id      object
cleaned_prenom     object
cleaned_nom        object
phoneticPrenom     object
phoneticNom        object
dtype: object
```

In [44]:

```
df22.dtypes
```

Out[44]:

```
A_name             object
A_nom              object
A_prenom           object
secondnames        object
cleaned_place      object
birth             float64
death             float64
Notes              object
cleaned_prenom     object
cleaned_nom        object
phoneticPrenom     object
phoneticNom        object
dtype: object
```

In [46]:

```
#essayer de joindre deux colonnes pour récupérer les noms et prénoms
#doit se faire en 2 étapes apparemment
result = pd.merge(matches3, df11[['B_name', 'birth', 'death']], left_on = ['belgium'],r
ight_index = True)
result.head(2)
```

Out[46]:

		N_exact	N_jarowink	P_exact	P_jarowink	P_Phon_exact	N_Phon_exact	d
belgium	**adieux**							
60	**28**	0	0.0	0	0.0	0	0	
	24	0	1.0	0	0.0	0	0	

In [48]:

```
#result3 = pd.merge(result2, dfadieux[['cleaned_name', 'cleaned_givenname']], left_on =
['lettres'],right_index = True)
result2 = pd.merge(result, df22[['A_name', 'birth','death']], left_on = ['adieux'],righ
t_index = True)
result2.head()
```

Out[48]:

		N_exact	N_jarowink	P_exact	P_jarowink	P_Phon_exact	N_Phon_exact	d
belgium	**adieux**							
60	**28**	0	0.0	0	0.0	0	0	
	24	0	1.0	0	0.0	0	0	
	23	0	1.0	0	0.0	0	0	
66	**66**	0	1.0	0	0.0	0	0	
	65	0	1.0	0	0.0	0	0	

In [49]:

```
result2.to_csv('data/potential_matches_lettresadieux.csv')
```

Annexe 5
Étude préliminaire des référentiels pour les noms de pays

Cette annexe présente les constats d'une étude préliminaire portant sur le choix d'un référentiel externe pour les noms de pays. Comme expliqué dans la sous-section 4.3.1 (page 190), ces entités géographiques n'ont pas encore été incorporées au sein de l'instance Wikibase développée dans le cadre de notre étude de cas – à l'exception des localités belges –, étant donné que ces choix se doivent d'être alignés avec la stratégie que les Archives de l'État choisiront d'adopter au cours des mois à venir. À court terme, seuls quelques pays limitrophes ont donc été ajoutés manuellement dans la Wikibase afin de couvrir les cas de figure rencontrés dans notre échantillon de données. Il nous semblait toutefois utile de présenter ici les types de questionnements entraînés par la sélection d'un tel référentiel, dans un contexte marqué par des corpus de données faisant parfois référence à des réalités aujourd'hui disparues.

Concernant les noms de pays, sachant que ni le CegeSoma ni les Archives ne disposaient d'une liste standardisée préexistante, la question s'est posée de la ressource à utiliser. Étant donné qu'il s'agit d'une liste de données, relativement peu soumise au changement, nous avonc décidé d'en créer une localement. Nous avons identifié trois sources potentielles :

1. les entités pays Wikidata
2. les codes pays de la norme ISO-3166
3. les codes pays du Service Public Fédéral (SPF) Affaires Étrangères

Premièrement, nous avons pensé aux pays décrits dans Wikidata. Cette première source d'information aurait le mérite de nous fournir des données pouvant être très facilement intégrées à notre Wikibase, en nous permettant d'obtenir d'un seul geste leur identifiant numérique, mais également des données complémentaires, telles qu'une description dans différentes langues, des formes alternatives du nom, d'autres identifiants (y compris le code ISO 3166-1), des dates de fondation ou encore la taille de la population. Étant donné qu'il s'agit de données reposant sur l'investissement de la communauté, l'élément critique ici est de savoir si la complétude et la qualité des informations est suffisante.

Pour obtenir cette liste de pays, trois types de requêtes ont été testées : la première demande que la nature de l'élément [P31] ait comme valeur : pays [Q6256] ; la seconde demande comme valeur : état souverain [Q671362] ; la troisième requête prolonge la seconde : elle cherche des états souverains

[Q671362] qui sont en outre décrits comme étant membres [P463] de l'Organisation des Nations Unies [Q1065].

La première requête [144] est basée sur la valeur *pays* [145]. À l'heure où nous écrivons [146], elle génère seulement 179 résultats. Si cette liste semble étonnament courte, c'est parce que certains éléments n'apparaissent pas, en raison des spécificités du modèle de données de Wikidata. En effet, le modèle permet de préciser un rang pour chaque valeur : rang normal, préféré ou obsolète [147] Or, si une valeur a été désignée comme rang préféré, les autres valeurs encodées pour cette propriété n'apparaitront plus (par défaut) dans les résultats de requêtes [148]. C'est pour cela que le Liechtenstein [Q347] n'apparaît pas dans nos 179 résultats : la valeur *État souverain* [Q3624078] a été désignée par un utilisateur de Wikidata comme rang préféré au détriment de *pays* [Q6256]. Sachant cela, nous pouvons remanier la requête pour inclure toutes les valeurs, indépendamment du rang qui leur a été assigné ; nous obtenons alors 230 résultats, incluant le Liechtenstein [149]. Cependant, quelle que soit la variante utilisée pour interroger Wikidata, tous les résultats ne semblent pas correspondre à nos besoins, qu'il s'agisse de données historiques remontant bien au-delà des besoins du CegeSoma, telles le Royaume d'Aksoum [Q139377] – un ancien État de la Corne de l'Afrique, dont le déclin remonte au XIe ou Xe siècle – ; ou de données dont la présence dans cette liste semble douteuse, tel l'Ordre souverain militaire hospitalier de Saint-Jean de Jérusalem, de Rhodes et de Malte' [Q190353], dont la description précise pourtant bien qu'il s'agit d'un *pseudo-État sans territoire*. En outre, certains éléments Wikidata n'apparaissent pas dans les résultats de cette requête, alors qu'ils auraient pourtant leur utilité dans le contexte du CegeSoma. C'est le cas d'anciens États, comme la Tchécoslovaquie [Q33946], la Yougoslavie [Q36704], la République démocratique allemande (RDA) [Q16957] ou encore l'Union des républiques sociales soviétiques (URSS) [Q15180].

La seconde requête [150] consiste à demander la liste des éléments ayant comme nature [P31] la valeur [Q3624078], c'est-à-dire un *état souverain* [151]. Cette seconde requête produit 203 résultats. À nouveau, la Yougoslavie et la Tchécoslovaquie n'y figurant pas, nous avons décidé de poursuivre l'expé-

144. https://w.wiki/FXm.
145. Décrit en français comme une « région généralement identifiée comme une entité géopolitique distincte ».
146. Le 15 janvier 2020.
147. Pour plus d'informations, voir https://www.wikidata.org/wiki/Help:Ranking/fr.
148. Voir https://www.wikidata.org/wiki/Wikidata_talk:SPARQL_query_service/queries#Query_not_returning_all_expected_results:_instance_of_country_not_returning_Cuba.
149. https://w.wiki/FYK.
150. https://w.wiki/FXn
151. Décrit en français comme une « organisation politique souveraine sur son territoire ».

rimentation en adaptant la requête pour prendre en compte toutes les valeurs indépendamment de leur rang [152]. Si nous relevons que la liste semble cette fois-ci beaucoup plus exhaustive (par exemple cinq occurences différentes comportent le mot Yougoslavie : Yougoslavie [Q36704] ; Royaume de Yougoslavie [Q191077] ; République fédérale de Yougoslavie [Q83861] ; République fédérative de Yougoslavie [Q83286] ; République fédérale de Yougoslavie [Q838261]), en revanche son étendue dépasse clairement nos besoins : des États historiques comme la Principauté de Tourov et Pinsk [Q671362] – une principauté médiévale située sur l'actuel territoire de la Biélorussie et de l'Ukraine – représenteraient davantage une source de bruit qu'une réelle plus-value au sein de notre instance Wikibase. Il serait imaginable de trier les États, par exemple en fonction de leurs dates d'existence, cette stratégie ne semble toutefois pas la plus heureuse : il s'agirait d'un processus laborieux sans garantie en ce qui concerne la qualité des résultats, ces derniers étant conditionnées par le degré de complétude des fiches Wikidata concernées.

En bref, ni la première ni la seconde requête ne semblent apporter de résultats satisfaisants. Une troisième requête [153] trouvée au cours de nos recherches pourrait toutefois s'avérer utile pour la suite : elle vise à obtenir tous les éléments Wikidata qui sont à la fois des États souverains et des membres actuels de l'Organisation des Nations unies (ONU). À défaut d'être exhaustive et de contenir des États n'existant plus aujourd'hui, elle a le mérite d'offrir une excellente base – incluant tous les codes ISO 3166 de ces pays – pour effectuer des renvois de la Wikibase vers Wikidata et pourra dès lors être utilisée pour la suite.

Deuxièmement, la seconde source envisagée est composée des codes pays de la norme ISO 3166-1. Cette norme internationale relative à la spécification de pays ET de régions d'intérêt général définit trois types de codes alphanumérique et numérique composés de deux ou trois caractères. Elle a pour but de définir « des codes internationalement reconnus de lettres et/ou de chiffres qui peuvent être utilisés pour désigner des pays et leurs subdivisions. Les noms des pays ne sont pas établis par l'ISO. Ils proviennent des listes des Nations Unies. » [154] Si cette liste promet une grande interopérabilité, en revanche elle ne comporte pas les codes pays devenus obsolètes suite à des fusions (comme l'Allemagne de l'Est et l'Allemagne de l'Ouest), scissions (la Tchécoslovaquie par exemple) ou changements de noms de pays (comme la Haute-Volta devenue Burkina Faso, par exemple). Pour cela, il faut se tourner vers la norme ISO 3166-3 qui inclut les codes pays à deux

152. https://w.wiki/FYb.
153. https://w.wiki/FYf.
154. https://www.iso.org/fr/iso-3166-country-codes.html

lettres ayant été retirés de la norme ISO 3166-1. Ces derniers sont complétés à l'aide du nouveau code (par exemple la Haute-Vola (HV), remplacée par le Burkina Faso devient HVBF) ou de code spécial non attribué lors de la scission d'un pays.

Une combinaison des codes pays des normes ISO 3166-1 et 3166-3 semble donc une solution raisonnable. Elle pourrait toutefois gagner à être combinée à la troisième source envisagée, qui est mise librement à disposition, contrairement aux codes pays ISO qui doivent être achetés.

La troisième ressource envisagée est la « Source authentique Codes pays », qui propose des codes INS destinés à être utilisés par l'Administration belge. Décrite comme le fruit d'un travail collaboratif entre le SPF Intérieur, la Banque Carrefour de la sécurité sociale, l'Office national de sécurité sociale, le SPF Économie, le SPF Finances, le SPF Justice et Fedict, cette Source comporte des « codes pays [...] qui ont valeur de données uniques et originales en Belgique, de sorte que d'autres instances ne doivent plus collecter ces codes » [155].

Cette ressource inclut des « pays », définis ici comme « le territoire d'un seul État », mais également des « régions dotées d'un code de représentation » qui ne sont pas forcément des pays, à l'instar de la Guadeloupe. L'Administration étant amenée à utiliser des pays aujourd'hui disparus, par exemple pour indiquer un lieu de naissance, cette ressource s'annonce théoriquement à même de répondre à nos besoins. Si cette liste est plus exhaustive que la norme ISO, elle inclut toutefois un alignement vers les codes ISO, estimant que dans le cadre d'échanges croissants de données électroniques « l'utilisation d'une norme internationale est un grand avantage ».

Avec cette ressource, ce n'est pas la question de l'exhaustivité mais celle de la granularité qui va se poser, dans la mesure où les codes INS ne distinguent par exemple pas le Danemark des Îles Féroé, qui possèdent tous deux le code 108, alors qu'ils comportent des codes ISO distincts – DK pour Danemark et FO pour Feroe Islands [156]. Nous pouvons donc nous poser la question de la plus-value de ces codes INS, s'ils ne possèdent pas d'identifiants uniques à un niveau de détail similaire à ceux proposés par la norme ISO. Par ailleurs, si cette source a l'avantage de fournir les formes courtes et les formes longues des noms de pays à la fois en français, en anglais et en néerlandais, il faut noter que la qualité des codes INS semble discutable dans la mesure où la Tchécoslovaquie apparaît par exemple deux fois, associée à des codes INS distincts – 130 et 171 –, sous des dates d'existence identiques

155. Voir : https://statbel.fgov.be/sites/default/files/Over_Statbel_FR/Nomenclaturen/NVcountrycodewhitpaper_fr.pdf.

156. https://statbel.fgov.be/sites/default/files/Over_Statbel_FR/Nomenclaturen/NVcountrycodewhitpaper_fr.pdf.

et associées au même code ISO – CS –, idem pour l'Allemagne de l'Est, qui possède deux codes INS distincts sans raison apparente.

Sans remplacer une analyse approfondie incluant par exemple la recherche de jeux de données directement disponibles au format RDF, ces quelques exemples donnent un aperçu des recherches et de l'arbitrage devant être réalisés afin de déterminer quelle source ou quelle combinaison de sources serait à même de répondre aux besoins de l'institution.

Annexe 6
Alignement des propriétés Wikibase avec les propriétés Wikidata et RiC-O

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P1	nature de l'élément	http://www.wikidata.org/entity/P31	nature de l'élément	https://www.ica.org/standards/RiC/ontology#belongsToCategory	belongs to category	La propriété rico:belongsToCategory, qui a pour portée Rico:Type, est utilisée si besoin est pour caractériser une ressource RiC-O lorsqu'on a déjà spécifié son appartenance à une classe RiC-O (comme Record ou Place) et qu'on veut employer pour cela un vocabulaire existant (comme un thésaurus des types de lieux ou des types de documents).
https://adochs.arch.be/entity/P2	identifiant Wikidata	/	/	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P3	élément(s) à combiner	/	/	/	/	/
https://adochs.arch.be/entity/P12	voir aussi	http://www.wikidata.org/entity/P1659	voir aussi	http://www.w3.org/2000/01/rdf-schema#seeAlso	/	RiC-O prévoit d'utiliser les propriétés définies nativement par OWL ou RDFS, ici rdfs:seeAlso.
https://adochs.arch.be/entity/P13	propriété(s) à combiner	/	/	/	/	/
https://adochs.arch.be/entity/P14	propriété équivalente	http://www.wikidata.org/entity/P1628	propriété équivalente	/	/	RiC-O ne contient pas encore d'équivalences OWL entre les propriétés RiC-O et des propriétés d'autres ontologies ; si c'était le cas, c'est la propriété owl:equivalentProperty qui serait utilisée.
https://adochs.arch.be/entity/P15	format de l'URL	https://www.wikidata.org/entity/P1630	format de l'url	/	/	/
https://adochs.arch.be/entity/P16	format de l'URI pour les ressources RDF	http://www.wikidata.org/entity/P1921	structure de l'URI pour les ressources RDF	/	/	/
https://adochs.arch.be/entity/P17	URL de la référence	http://www.wikidata.org/entity/P854	URL de la référence	https://www.ica.org/standards/RiC/ontology#hasSource	has source	Une équivalence peut être établie avec rico:hasSource, bien que cette propriété soit plus spécifique : elle permet de donner l'URI d'un RecordResource ou d'un Agent qui est la source utilisée pour établir une relation, ou la source d'une ressource archivistique.
https://adochs.arch.be/entity/P18	date de consultation	http://www.wikidata.org/entity/P813	date de consultation	https://www.ica.org/standards/RiC/ontology.html#isDateAssociatedWith	is associated with date	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:isAssociatedWithDate est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:isAssociatedWithDate.

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P19	identifiant Pallas	/	/	https://www.ica.org/standards/RiC/ontology#identifier	identifier	L'équivalence n'est pas totale dans la mesure où il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:identifier.
https://adochs.arch.be/entity/P20	coordonnées géographiques	http://www.wikidata.org/entity/P625	coordonnées géographiques	https://www.ica.org/standards/RiC/ontology#geographicalCoordinates	geographical coordinates	À noter : rico:geographicalCoordinates "may be deprecated and removed later on. Use only if you don't use PhysicalLocation and Coordinates classes with Place."
https://adochs.arch.be/entity/P21	date	http://www.wikidata.org/entity/P585	date	https://www.ica.org/standards/RiC/ontology#Date	date	La définition de la propriété rico:Date (qui associe une date à n'importe quelle chose (rico:Thing)) est plus large car elle ne se concentre pas sur un événement ou une déclaration, comme le fait la propriété Wikidata P585.
https://adochs.arch.be/entity/P22	date de début	http://www.wikidata.org/entity/P580	date de début	https://www.ica.org/standards/RiC/ontology#beginningDate	beginning date	La définition de la propriété rico:beginningDate est plus large car elle ne se concentre par sur une déclaration, par ailleurs il faut que la propriété "may be deprecated and removed later on".
https://adochs.arch.be/entity/P23	date de fin	http://www.wikidata.org/entity/P582	date de fin	https://www.ica.org/standards/RiC/ontology#endDate	end date	La définition de la propriété rico:endDate est plus large car elle ne se concentre par sur une déclaration, par ailleurs il faut que la propriété "may be deprecated and removed later on".
https://adochs.arch.be/entity/P24	partie de	http://www.wikidata.org/entity/P361	partie de	https://www.ica.org/standards/RiC/ontology#isPartOf	is part of	
https://adochs.arch.be/entity/P25	qualité de l'information	http://www.wikidata.org/entity/P1480	qualité de l'information	https://www.ica.org/standards/RiC/ontology#certainty	certainty	L'équivalence est limitée dans la mesure où la définition de la propriété rico:certainty est plus restreinte : elle définit le degré de certitude d'une date, d'un événement ou d'une relation ; la propriété Wikidata P1480 a un périmètre plus vaste dans la mesure où elle concerne également la certitude et la précision.
https://adochs.arch.be/entity/P28	titre de noblesse	http://www.wikidata.org/entity/P97	titre de noblesse	https://www.ica.org/standards/RiC/ontology#hasAgentName	has agent name	L'équivalence est limitée dans la mesure où la propriété rico:hasAgentName est plus large (elle a comme portée agent: AgentName, qui est définie ainsi : "a label, title or term designating an Agent in order to make it distinguishable from other similar entities"), il s'agirait ici d'étendre RiC-O en définissant une sous-propriété.
https://adochs.arch.be/entity/P29	alias	http://www.wikidata.org/entity/P742	pseudonyme	https://www.ica.org/standards/RiC/ontology#name	name	L'équivalence est limitée dans la mesure où la propriété rico:name est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:name.
https://adochs.arch.be/entity/P31	date de naissance	http://www.wikidata.org/entity/P569	date de naissance	https://www.ica.org/standards/RiC/ontology#hasBirthDate	has birth date	

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P32	date de décès	http://www.wikidata.org/entity/P570	date de mort	https://www.ica.org/standards/RiC/ontology.html#hasDeathDate	has death date	
https://adochs.arch.be/entity/P33	lieu de naissance	http://www.wikidata.org/entity/P19	lieu de naissance	/	/	Il n'existe pas d'équivalence directe (Ric-R325, correspondant à 'had birth place' dans Ric-cm-01, n'est pas repris dans Ric-cm-02 et RiC-O), il faudrait donc lier la personne à un événement via 'rico:affectedBy', cet événement pouvant être qualifié de lieu de naissance via l'attribut Event Type.
https://adochs.arch.be/entity/P34	lieu de décès	http://www.wikidata.org/entity/P20	lieu de mort	/	/	Il n'existe pas d'équivalence directe (Ric-R326, correspondant à 'had death place' dans Ric-cm-01, n'est pas repris dans Ric-cm-02 et RiC-O), il faudrait donc lier la personne à un événement via 'rico:affectedBy', cet événement pouvant être qualifié de lieu de naissance via l'attribut Event Type.
https://adochs.arch.be/entity/P35	circonstances de la mort	http://www.wikidata.org/entity/P1196	circonstances de la mort	/	/	Dans RiC-O, la mort d'une personne peut être définie comme une instance d'un événement (rico:Event) affectant une personne, qui peut avoir une date, un lieu, une description, etc. Un événement peut aussi résulter de (rico:resultsFrom) un autre événement, ou être la cause (rico:resultsIn) d'un autre événement ou d'une instance de rico:Thing ; et impliquer (rico:involves) n'importe quelle chose.
https://adochs.arch.be/entity/P36	cause de la mort	http://www.wikidata.org/entity/P509	cause de la mort	/	/	Voir ligne précédente.
https://adochs.arch.be/entity/P37	genre	http://www.wikidata.org/entity/P21	sexe ou genre	/	/	Il n'existe pas d'équivalence directe, en revanche il existe une propriété rico:hasDemographicGroup, avec pour range la classe rico:DemographicGroup, dont on pourrait définir des sous-classes liées au genre.
https://adochs.arch.be/entity/P38	pays de nationalité	http://www.wikidata.org/entity/P27	pays de nationalité	https://www.ica.org/standards/RiC/ontology#isAssociatedWithPlace	is associated with place	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:isAssociatedWithPlace est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété pour les pays de nationalités.
https://adochs.arch.be/entity/P39	langue(s)	http://www.wikidata.org/entity/P1412	langues parlées, écrites ou signées	https://www.ica.org/standards/RiC/ontology#hasLanguage	has language	L'équivalence n'est pas complète dans la mesure où la propriété rico:hasLanguage s'applique tant à une personne qu'à une ressource.
https://adochs.arch.be/entity/P40	résidence	http://www.wikidata.org/entity/P551	résidence	https://www.ica.org/standards/RiC/ontology#hasLocation	has location	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:hasLocation est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété pour le lieu de résidence d'une personne.

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P41	parentèle	http://www.wikidata.org/entity/P1038	parentèle	https://www.ica.org/standards/RiC/ontology#hasFamilyLinkWith	has family link with	
https://adochs.arch.be/entity/P42	nombre d'enfants	http://www.wikidata.org/entity/P1971	nombre d'enfants	/	/	/
https://adochs.arch.be/entity/P43	sœur ou frère	http://www.wikidata.org/entity/P3373	frère ou sœur	https://www.ica.org/standards/RiC/ontology.html#hasSibling	has sibling	
https://adochs.arch.be/entity/P44	père	http://www.wikidata.org/entity/P3373	père	https://www.ica.org/standards/RiC/ontology#hasParent	has parent	L'équivalence n'est pas parfaite dans la mesure ou rico:hasParent est plus générique et ne distingue pas le père de la mère, mais une sous-propriété pourrait être créée pour le faire.
https://adochs.arch.be/entity/P45	mère	http://www.wikidata.org/entity/P25	mère	https://www.ica.org/standards/RiC/ontology#hasParent	has parent	L'équivalence n'est pas parfaite dans la mesure ou rico:hasParent est plus générique et ne distingue pas le père de la mère, mais une sous-propriété pourrait être créée pour le faire.
https://adochs.arch.be/entity/P46	conjoint·e	http://www.wikidata.org/entity/P26	conjoint	https://www.ica.org/standards/RiC/ontology#hasSpouse	has spouse	
https://adochs.arch.be/entity/P47	identifiant ISNI	http://www.wikidata.org/entity/P213	identifiant ISNI	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P48	identifiant VIAF	http://www.wikidata.org/entity/P214	identifiant VIAF	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P49	identifiant SNAC	http://www.wikidata.org/entity/P3430	identifiant Social Networks Archival Context	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P50	identifiant EHRI	/	/	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P51	identifiant GeoNames	http://www.wikidata.org/entity/P1566	identifiant GeoNames	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P52	identifiant ODIS	http://www.wikidata.org/entity/P2372	identifiant Odis	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P53	code INS	http://www.wikidata.org/entity/P1567	code INS	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P54	occupation	http://www.wikidata.org/entity/P106	occupation	https://www.ica.org/standards/RiC/ontology#hasOccupationOfType	has occupation of type	
https://adochs.arch.be/entity/P55	parti politique	http://www.wikidata.org/entity/P102	parti politique	https://www.ica.org/standards/RiC/ontology#isMemberOf	is member of	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:isMemberOf est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété pour l'appartenance à un parti politique.
https://adochs.arch.be/entity/P56	image	http://www.wikidata.org/entity/P18	image	/	/	/
https://adochs.arch.be/entity/P57	identifiant AGR	/	/	https://www.ica.org/standards/RiC/ontology#Identifier	identifier	L'équivalence n'est pas totale, dans la mesure où la propriété rico:identifier est générique, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété comme rico:wikidataIdentifier.
https://adochs.arch.be/entity/P58	localisation administrative	http://www.wikidata.org/entity/P131	localisation administrative	https://www.ica.org/standards/RiC/ontology#hasLocation	has location	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:hasLocation est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété propre à la dimension administrative.
https://adochs.arch.be/entity/P59	issu de	/	/	https://www.ica.org/standards/RiC/ontology#hasSource	has source	
https://adochs.arch.be/entity/P60	importé de Wikidata	/	/	https://www.ica.org/standards/RiC/ontology#hasSource	has source	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:hasSource est plus large.
https://adochs.arch.be/entity/P61	personnalités Belgium WWII	/	/	https://www.ica.org/standards/RiC/ontology#describedBy	described by	Cette équivalence peut être établie avec la propriété plus générique rico:describedBy si l'on considère la page web concernée comme un document d'archives.
https://adochs.arch.be/entity/P62	conflit	http://www.wikidata.org/entity/P607	conflit	https://www.ica.org/standards/RiC/ontology#involvedIn	involved in	Cette équivalence peut être établie avec la propriété plus générique rico:involvedIn
https://adochs.arch.be/entity/P66	prénom	http://www.wikidata.org/entity/P735	prénom	https://www.ica.org/standards/RiC/ontology#name	name	L'équivalence est limitée dans la mesure où la propriété rico:name est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:name pour les prénoms.
https://adochs.arch.be/entity/P67	nom de famille	http://www.wikidata.org/entity/P734	nom de famille	https://www.ica.org/standards/RiC/ontology#name	name	L'équivalence est limitée dans la mesure où la propriété rico:name est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:name pour les noms de famille.

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P68	distinction reçue	http://www.wikidata.org/entity/P166	distinction reçue	/	/	/
https://adochs.arch.be/entity/P69	sujet de			https://www.ica.org/standards/RiC/ontology.html#isSubjectOf	is subject of	La propriété rico:hasMainSubject pourrait également être utilisée.
https://adochs.arch.be/entity/P70	nom de naissance	http://www.wikidata.org/entity/P1477	nom de naissance	https://www.ica.org/standards/RiC/ontology#name	name	L'équivalence est limitée dans la mesure où la propriété rico:name est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété de rico:name pour les prénoms de naissance.
https://adochs.arch.be/entity/P71	état civil	/	/	/	/	/
https://adochs.arch.be/entity/P72	producteur de	/	/	https://www.ica.org/standards/RiC/ontology#isProvenanceOf	is provenance of	À noter que RiC-O permet une plus grande précision, par exemple en utilisant les propriétés rico:isCreatorOf ou rico:accumulates.
https://adochs.arch.be/entity/P73	code ISO 3166-1 alpha-2 du pays	http://www.wikidata.org/entity/P297	code ISO 3166-1 alpha-2 du pays	https://www.ica.org/standards/RiC/ontology#identifiedBy	identified by	
https://adochs.arch.be/entity/P74	membre de	http://www.wikidata.org/entity/P463	membre de	https://www.ica.org/standards/RiC/ontology#isMemberOf	is member of	
https://adochs.arch.be/entity/P75	type de résistance	/	/	/	/	/
https://adochs.arch.be/entity/P76	fonction	http://www.wikidata.org/entity/P39	fonction	https://www.ica.org/standards/RiC/ontology#occupies	occupies	
https://adochs.arch.be/entity/P77	demande de statut(s) de reconnaissance nationale	/	/	/	/	/
https://adochs.arch.be/entity/P78	lieu de détention	http://www.wikidata.org/entity/P2632	lieu de détention	https://www.ica.org/standards/RiC/ontology#isAssociatedWithPlace	is associated with place	L'équivalence n'est pas totale dans la mesure où la définition de la propriété rico:isAssociatedWithPlace est plus large, il s'agirait ici d'étendre RiC-O en définissant une sous-propriété pour les lieux de détention. Il serait également possible de créer un événement pour la détention (rico:affectedBy + rico:Event (la détention) ; et depuis rico:Event, rico:hasLocation + rico:Place).
https://adochs.arch.be/entity/P79	événement clé	http://www.wikidata.org/entity/P793	événement clé	https://www.ica.org/standards/RiC/ontology#affectedBy	affected by	

Wikibase (Adochs)	WB_rdfs: label@fr	Wikidata	WD_rdfs: label@fr	RiC-O	RICO_rdfs: label@en	Commentaires concernant RiC-O (Florence Clavaud & Anne Chardonnens)
https://adochs.arch.be/entity/P80	condamné·e pour	http://www.wikidata.org/entity/P1399	condamné pour	/	/	L'équivalence n'est pas directe, il s'agirait de créer et décrire un événement lié à la condamnation : rico:affectedBy + rico:Event (condamnation) ; rico:Event rico:descriptiveNote.
https://adochs.arch.be/entity/P81	motif détaillé	/	/	/	/	/
https://adochs.arch.be/entity/P82	condamnation	http://www.wikidata.org/entity/P1596	condamnation	/	/	L'équivalence n'est pas directe, il s'agirait de créer et décrire un événement lié à la condamnation : rico:affectedBy + rico:Event (condamnation) ; rico:Event rico:descriptiveNote.
https://adochs.arch.be/entity/P83	requête en grâce	/	/	/	/	/
https://adochs.arch.be/entity/P84	note descriptive	/	/	https://www.ica.org/standards/RiC/ontology#descriptiveNote	descriptive note	
https://adochs.arch.be/entity/P85	affirmé dans	http://www.wikidata.org/entity/P248	affirmé dans	https://www.ica.org/standards/RiC/ontology#source	source	Si c'est la valeur est une chaîne de caractères comme dans le cas de Wikibase, c'est rico:source qui doit être utilisé, dans le cadre de la propriété Wikidata P248, la propriété équivalente serait plus rico:hasSource.
https://adochs.arch.be/entity/P86	sous-classe de	http://www.wikidata.org/entity/P279	sous-classe de	https://www.w3.org/2000/01/rdf-schema#subClassOf	subClassOf	RiC-O prévoit d'utiliser les propriétés définies nativement par OWL ou RDFS, ici rdfs:subClassOf.
Libellé Wikibase (rdfs:label)	libellé	rdfs:label	libellé	https://www.w3.org/2000/01/rdf-schema#label		RiC-O prévoit d'utiliser les propriétés définies nativement par OWL ou RDFS, ici rdfs:label.
Description Wikibase (schema:description)	description	schema:description	description	https://www.ica.org/standards/RiC/ontology#descriptiveNote	descriptive note	

Annexe 7
Paramètres d'installation de Wikibase

Ces détails de l'installation de l'instance Wikibase sur les serveurs des Archives générales du Royaume nous ont été communiqués par email le 31 mars 2020, par Peter Van Overveldt, responsable ICT.

1. A virtual machine *adochsdockerVM* was created on our Virtual-machine-host *sumi0002*. *sumi0002* is a « Proxmox »-installation ; the underying virtualisation technology is « KVM ».

   Highlights of the *adochsdockerVM*-machine :
   — 8GB memory
   — 2 virtual processors, which on this host operate at 3,6GhZ each
   — Installed operating system : « Linux Ubuntu 18.4 LTS (Long term Support) server edition »
   — The network connects to the so called DMZ-network of the State Archives.

2. Docker and docker-compose were installed on *adochsdockerVM*, in order to be able to load the docker-version of Wibkibase [sic] on it.

   Remark : at the moment, the State Archives do not yet have fully deployed Docker-platform like *Kubernetes* or « Docker Swarm ». So, for THIS INSTALLATION, only the containers regarding this Wikibase run on the adochsdocker Virtual-Machine. Doing so, we take advantage of the ease-of-installation and -maintenance the *docker*-technology offers. Platforms like *Kubernetes* or « Docker Swarm » would offer additional advantages like better resource sharing or resilience against hardware failure, but are also hard(er) to deploy.

3. *Wikibase* was *pulled* onto the *adochsdocker*-VM, making use of an adapted version of file https://raw.githubusercontent.com/wmde/wikibase-docker/master/docker-compose.yml, and command *docker-compose*.

4. Deploying the wikibase-installation on the internet (and, in fact, the internal network as well), i.e. :
   — These names were configured with our internet DNS provider (Belnet) :
      (a) adochs.arch.be
      (b) quickstatements-adochs.arch.be
      (c) query-adochs.arch.be

— SSL Certificates were requested and obtained for these names from our « Certificate Authority » *Digicert*.
— The reverse proxy (an *HAProxy*-installation) of the State Archives, which already existed prior to the Adochs-project, was additionally configured to :
  (a) forward requests on https://adochs.arch.be to http://"adochsdockerIPaddress":80
  (b) forward requests on https://quickstatement-adochs.arch.be to http://"adochsdockerIPaddress":9191
  (c) forward requests on https://query-adochs.arch.be to http://"adochsdockerIPaddress":8282
— *Remarks*
  — So, as you can see, the reverse proxy is also performing so called SSL-offloading (i.e. : it does the encrypting and decrypting for the "https"-protocol on behalf of the actual adochs-server)
  — For improving security, the reverse proxy was also configured to prevent old encryption technologies like TLS 1.0, TLS 1.1 or SSLv3 to be used by clients.

Annexe 8
Configuration de la Wikibase

Pour plus d'infos au sujet de ce document, consulter cette page (https://linkingthepast.org/about/).

Configuration Wikibase

De nombreuses sources* de documentation existent déjà au sujet instance Wikibase. Le présent document vise donc uniquement à partager les paramètres de configuration utilisés dans le cadre de notre prototype 'DataCegesoma'.

- Voici un récapitulatif aussi exhaustif que possible de ces sources :
- Wikibase (https://wikiba.se/)
- LearningWikibase (http://learningwikibase.com)
- Phabricator, Wikibase-Containers (https://phabricator.wikimedia.org/project/profile/3079/)
- Wikibase Community User Group (https://meta.wikimedia.org/wiki/Wikibase_Community_User_Group)
- Telegram group (https://t.me/joinchat/HGjGexZ9NE7BwpXzMsoDLA)
- Wikibase Community User Group - \emph{Mailing list} (https://lists.wikimedia.org/mailman/listinfo/wikibaseug)
- MediaWiki - Wikibase FAQ (https://www.mediawiki.org/wiki/Wikibase/FAQ)
- MediaWiki - Wikibase Installation (https://www.mediawiki.org/wiki/Wikibase/Installation)
- Contact the development team (https://www.wikidata.org/wiki/Wikidata:Contact_the_development_team)
- Using OpenStack to run a custom Wikibase (https://fuga.cloud/labs/using-openstack-to-run-custom-wikibase/)
- Wikibase for Research Infrastructure — Part 1 (https://link.medium.com/wH6OkVlvJ6)
- Wikibase Install Basic Tutorial (https://semlab.io/howto/wikibase_basic)
- Running and querying my own Wikibase instance (http://www.snee.com/bobdc.blog/2018/06/running-and-querying-my-own-wi.html)
- Wikibase: configure, customize, and collaborate (https://stuff.coffeecode.net/2018/wikibase-workshop-swib18.html)
- Installing Blazegraph and Wikibase (https://heardlibrary.github.io/digital-scholarship/lod/install/)
- 2 minutes on installing Wikibase (https://youtu.be/P174BEDhUJg)
- Meta-Wiki - Wikibase Upgrade Workflow (https://meta.wikimedia.org/wiki/File:Wikibase_Upgrade_Workflow.pdf)
- Wikidata - Wikibase documentation (https://www.wikidata.org/wiki/Wikidata:Wikibase_documentation)
- GitHub repository: wikibase-docker (https://github.com/wmde/wikibase-docker/blob/master/README.md)
- Addhsore's blog posts (https://addshore.com/tag/wikibase/)

Fichier *docker-compose.yml*

In [ ]:

```
# Wikibase with Query Service
#
# This docker-compose example can be used to pull the images from docker hub.
#
# Examples:
#
# Access Wikibase via "http://localhost:8181"
#   (or "http://$(docker-machine ip):8181" if using docker-machine)
#
# Access Query Service via "http://localhost:8282"
#   (or "http://$(docker-machine ip):8282" if using docker-machine)
version: '3'

services:
  wikibase:
    image: wikibase/wikibase:1.33-bundle #wikibase/wikibase:1.33-bundle
    links:
      - mysql
    ports:
    # CONFIG - Change the 8181 here to expose Wikibase & MediaWiki on a different port
      - "80:80" #"8181:80"
    volumes:
      - mediawiki-images-data:/var/www/html/images
      - quickstatements-data:/quickstatements/data
    depends_on:
    - mysql
    - elasticsearch
    restart: unless-stopped
    networks:
      default:
        aliases:
         - wikibase.svc
           #- adochs1.arch.be
           # CONFIG - Add your real wikibase hostname here, for example wikibase-regist
ry.wmflabs.org
    environment:
      - DB_SERVER=mysql.svc:3306
      - MW_ELASTIC_HOST=elasticsearch.svc
      - MW_ELASTIC_PORT=9200
      # CONFIG - Change the default values below
      - MW_ADMIN_NAME=*** #WikibaseAdmin
      - MW_ADMIN_PASS=*** #WikibaseDockerAdminPass
      - MW_ADMIN_EMAIL=*** #admin@example.com
      - MW_WG_SECRET_KEY=secretkey
      # CONFIG - Change the default values below (should match mysql values in this fil
e)
      - DB_USER=wikiuser
      - DB_PASS=sqlpass
      - DB_NAME=my_wiki
      - QS_PUBLIC_SCHEME_HOST_AND_PORT=https://quickstatements-adochs.arch.be:443
  mysql:
    image: mariadb:10.3
    restart: unless-stopped
    volumes:
      - mediawiki-mysql-data:/var/lib/mysql
    environment:
```

```
      MYSQL_RANDOM_ROOT_PASSWORD: 'yes'
      # CONFIG - Change the default values below (should match values passed to wikibas
e)
      MYSQL_DATABASE: 'my_wiki'
      MYSQL_USER: 'wikiuser'
      MYSQL_PASSWORD: 'sqlpass'
    networks:
      default:
        aliases:
         - mysql.svc
  wdqs-frontend:
    image: wikibase/wdqs-frontend:latest
    restart: unless-stopped
    ports:
    # CONFIG - Change the 8282 here to expose the Query Service UI on a different port
     - "8282:80"
    depends_on:
    - wdqs-proxy
    networks:
      default:
        aliases:
         - wdqs-frontend.svc
    environment:
      - WIKIBASE_HOST=adochs.arch.be #wikibase.svc
      - WDQS_HOST=wdqs-proxy.svc
      - BRAND_TITLE=DataCegeSoma Query Service
  wdqs:
    image: wikibase/wdqs:0.3.10
    restart: unless-stopped
    volumes:
      - query-service-data:/wdqs/data
    command: /runBlazegraph.sh
    networks:
      default:
        aliases:
         - wdqs.svc
    environment:
      - WIKIBASE_HOST=adochs.arch.be #wikibase.svc
      - WIKIBASE_SCHEME=https
      - WDQS_HOST=wdqs.svc
      - WDQS_PORT=9999
    expose:
      - 9999
  wdqs-proxy:
    image: wikibase/wdqs-proxy
    restart: unless-stopped
    environment:
      - PROXY_PASS_HOST=wdqs.svc:9999
    ports:
     - "8989:80"
    depends_on:
    - wdqs
    networks:
      default:
        aliases:
         - wdqs-proxy.svc
  wdqs-updater:
    image: wikibase/wdqs:0.3.10
    restart: unless-stopped
```

```
    command: /runUpdate.sh
    depends_on:
    - wdqs
    - wikibase
    networks:
      default:
        aliases:
         - wdqs-updater.svc
    environment:
     - WIKIBASE_HOST=adochs.arch.be #wikibase.svc
     - WIKIBASE_SCHEME=https
     - WDQS_HOST=wdqs.svc
     - WDQS_PORT=9999
  elasticsearch:
    image: wikibase/elasticsearch:5.6.14-extra
    restart: unless-stopped
    networks:
      default:
        aliases:
         - elasticsearch.svc
    environment:
      discovery.type: single-node
      ES_JAVA_OPTS: "-Xms512m -Xmx512m"
  # CONFING, in order to not load quickstatements then remove this entire section
  quickstatements:
    image: wikibase/quickstatements:latest
    ports:
     - "9191:80"
    depends_on:
     - wikibase
    volumes:
     - quickstatements-data:/quickstatements/data
    networks:
      default:
        aliases:
         - quickstatements.svc
           #- adochs1.arch.be
           # - adochs.arch.be
    environment:
     - QS_PUBLIC_SCHEME_HOST_AND_PORT=https://quickstatements-adochs.arch.be:443  #htt
p://localhost:9191
     - WB_PUBLIC_SCHEME_HOST_AND_PORT=https://adochs.arch.be:443  #http://localhost:818
1
     - WIKIBASE_SCHEME_AND_HOST=https://adochs.arch.be #http://wikibase.svc
     - WB_PROPERTY_NAMESPACE=122
     - "WB_PROPERTY_PREFIX=Property:"
     - WB_ITEM_NAMESPACE=120
     - "WB_ITEM_PREFIX=Item:"

volumes:
  mediawiki-mysql-data:
  mediawiki-images-data:
  query-service-data:
  quickstatements-data:
```

Fichier *localsettings.php*

In [ ]:

```
<?php
/**
 * ------------------------------------------------------------------------------------
------
 * This file is provided by the wikibase/wikibase docker image.
 * This file will be passed through envsubst which will replace "$" with "$".
 * If you want to change MediaWiki or Wikibase settings then either mount a file over
this
 * template and or run a different entrypoint.
 * ------------------------------------------------------------------------------------
------
 */

## Database settings
## Environment variables will be substituted in here.
$wgDBserver = "mysql.svc:3306";
$wgDBname = "my_wiki";
$wgDBuser = "wikiuser";
$wgDBpassword = "sqlpass";

## Logs
## Save these logs inside the container
$wgDebugLogGroups = [
        'resourceloader' => '/var/log/mediawiki/resourceloader.log',
        'exception' => '/var/log/mediawiki/exception.log',
        'error' => '/var/log/mediawiki/error.log',
];

## Site Settings
# TODO pass in the rest of this with env vars?
$wgShellLocale = "en_US.utf8";
$wgLanguageCode = "${MW_SITE_LANG}"; #MODIF, cf. ec-doris
$wgSitename = "DataCegeSoma"; #MODIF
$wgMetaNamespace = "Project";
# Configured web paths & short URLs
# This allows use of the /wiki/* path
## https://www.mediawiki.org/wiki/Manual:Short_URL
$wgScriptPath = "/w";        // this should already have been configured this way
#$wgScriptPath= ""; #cf. Seb35
$wgArticlePath = "/wiki/$1";

#Set Secret
$wgSecretKey = "secretkey";

## RC Age
# https://www.mediawiki.org/wiki/Manual:
# Items in the recentchanges table are periodically purged; entries older than this m
any seconds will go.
# The query service (by default) loads data from recent changes
# Set this to 1 year to avoid any changes being removed from the RC table over a shor
ter period of time.
$wgRCMaxAge = 365 * 24 * 3600;

wfLoadSkin( 'Vector' );

## Wikibase
```

```
# Load Wikibase repo, client & lib with the example / default settings.
require_once "$IP/extensions/Wikibase/lib/WikibaseLib.php";
require_once "$IP/extensions/Wikibase/repo/Wikibase.php";
require_once "$IP/extensions/Wikibase/repo/ExampleSettings.php";
require_once "$IP/extensions/Wikibase/client/WikibaseClient.php";
require_once "$IP/extensions/Wikibase/client/ExampleSettings.php";

# OAuth
wfLoadExtension( 'OAuth' );
$wgGroupPermissions['sysop']['mwoauthproposeconsumer'] = true;
$wgGroupPermissions['sysop']['mwoauthmanageconsumer'] = true;
$wgGroupPermissions['sysop']['mwoauthviewprivate'] = true;
$wgGroupPermissions['sysop']['mwoauthupdateownconsumer'] = true;

# WikibaseImport
require_once "$IP/extensions/WikibaseImport/WikibaseImport.php";

# CirrusSearch
wfLoadExtension( 'Elastica' );
require_once "$IP/extensions/CirrusSearch/CirrusSearch.php";
$wgCirrusSearchServers = [ 'elasticsearch.svc' ];
$wgSearchType = 'CirrusSearch';
$wgCirrusSearchExtraIndexSettings['index.mapping.total_fields.limit'] = 5000;

# UniversalLanguageSelector
wfLoadExtension( 'UniversalLanguageSelector' );

# cldr
wfLoadExtension( 'cldr' );

# EntitySchema
wfLoadExtension( 'EntitySchema' );



######
#
#     AJOUTS ANNE (03/04/2020)
#
#####

# add a logo
$wgLogo = "$wgResourceBasePath/images/logo.jpeg";


# Cf. installation Seb35 - NOEMI
#unset ( $wgExtraNamespaces[WB_NS_ITEM] );
#unset ( $wgExtraNamespaces[WB_NS_ITEM_TALK] );
#$wgWBRepoSettings['entityNamespaces']['item'] = NS_MAIN;

$wgWBRepoSettings['siteLinkGroups']=[]; #we don't need sitelinks

$wgLanguageSelectorLanguages= array ("fr","nl","en","de");

# droits et acces

#disallow anonymous editing
$wgGroupPermissions['*']['edit']=false;
```

```
#don't let users create their own accounts
$wgGroupPermissions['*']['createaccount']=true;
$wgGroupPermissions['sysop']['createaccount']= true;


# Configure formatterUrl and CanonicalURI
# $wgWBRepoSettings['formatterURLProperty']='P';
# $wgWBRepoSettings['canonicalUriProperty']='P';

# Add this to separate the identifiers in a separate section
$wgWBRepoSettings['statementSection']= array(
        'item' => array(
                'statements' => null,
                'identifiers' => array(
                        'type' => 'dataType',
                        'dataTypes' => array( 'external-id' ),
                ),

        ),
);


#Allowing file upload
$wgEnableUploads = true;



## EXTENSIONS
#
wfLoadExtension( 'Gadgets' );

#wfLoadExtension( 'VisualEditor' ); pas compatible avec mediawiki 1.33

wfLoadExtension( 'WikiEditor' );
$wgHiddenPrefs[] = 'usebetatoolbar';

wfLoadExtension('CodeEditor');


#wfLoadExtension('MobileFrontend');
#$wgMFAutodetectMobileView = true;
#$wgMFDefaultSkinClass = 'SkinVector';

wfLoadExtension( 'ConfirmEdit');
wfLoadExtensions(['ConfirmEdit','ConfirmEdit/ReCaptchaNoCaptcha']);

$wgCaptchaClass = 'ReCaptchaNoCaptcha';
$wgReCaptchaSiteKey = '6LfiJegUAAAAAHXtXG4ALYuv0roV4Qo2cwG4kVlb';
$wgReCaptchaSecret = '6LfiJegUAAAAAD6RgOWwmHx2dmxx5Ub7UMULA6eX';

$wgMainCacheType = CACHE_ANYTHING;

$wgCaptchaTriggers['edit'] =false;
$wgCaptchaTriggers['create']=false;
$wgCaptchaTriggers['createtalk'] =true;
$wgCaptchaTriggers['addurl']=false;
$wgCaptchaTriggers['createaccount'] = true;
$wgCaptchaTriggers['badlogin']= true;
$wgCaptchaTriggers['login']=true;
```

```
$wgGroupPermissions['*']['skipcaptcha']=false;
$wgGroupPermissions['user']['skipcaptcha']=false;
$wgGroupPermissions['autoconfirmed']['skipcaptcha']=false;
$wgGroupPermissions['bot']['skipcaptcha']=true; #registered bots
$wgGroupPermissions['sysop']['skipcaptcha']=true;
$wgGroupPermissions['editors']['skipcaptcha']=true;
```

Annexe 9
Interrogation de la Wikibase

Pour accéder à ce Jupyter Notebook au format .ipynb, suivre ce lien (...). Pour plus d'infos, consulter cette page (https://linkingthepast.org/about/).

Interroger sa propre instance Wikibase en Python

Cette étape vise à explorer comment nous pouvons interroger / manipuler / enregistrer les données contenues dans une instance Wikibase en faisant appel à un Jupyter Notebook et aux librairies Python *Requests*, *Pandas* et *SPARQLWrapper*.

Objectif

Le but ici est de tester les deux façons dont nous pouvons utiliser le langage Python pour interroger la Wikibase sans devoir passer par l'interface graphique :

- 1/ En passant par l'**API MediaWiki**
- 2/ En passant par le **SPARQL Endpoint**

1. MediaWiki API

Inspiration : Wikidata Training: the Mediawiki API (https://tools.wmflabs.org/paws-public/30793854/Wikidata%20Mediawiki%20API.ipynb)

Cette méthode offre des possibilités plus limitées que le SPARQL endpoint (qui permet de formuler des requêtes plus complexes, après la conversion des données en RDF) et son usage est soumis à des limites*, mais les modules *wbgetentities* et *wbsearchentities* permettent toutefois d'accéder facilement au "JSON canonique des pages d'entités". Cela peut être utile par exemple si l'on souhaite obtenir toutes les informations disponibles sur une entité en particulier.

*Voir la documentation de l'API (https://www.mediawiki.org/wiki/API:Main_page).

In [1]:

```
import requests
import pandas as pd

SSL_VERIFY = True
# maybe set SSL_VERIFY to False if connection to https://www.wikidata.org doesn't work
 (e.g. because of a proxy)
# To disable the SSL verification, remove comment sign (#) from next line
#SSL_VERIFY = False
#if not SSL_VERIFY:
#    import urllib3
#    urllib3.disable_warnings()
```

1.1 Obtenir les données d'entités Wikibase

Dans cet exemple, nous lançons un appel à l'API à l'aide de l'action **wbgetentities** afin d'obtenir toutes les données concernant Andrée de Jongh possédant l'identifiant [Q10] (https://adochs.arch.be/w/index.php?title=Item:Q10 (https://adochs.arch.be/w/index.php?title=Item:Q10)), en français. Sachez toutefois qu'il est possible d'obtenir les données de plusieurs entités à la fois (jusqu'à 50). Il est également possible de récupérer les données dans toutes les langues disponibles (il suffit de ne rien préciser).

In [2]:

```
get_dejongh = 'https://adochs.arch.be/w/api.php?action=wbgetentities&ids=Q10&format=jso
n&languages=fr'
res = requests.get(get_dejongh, verify=SSL_VERIFY)
result1 = res.json()
```

In [3]:

```
print(result1)
```

```
{'entities': {'Q10': {'pageid': 13, 'ns': 120, 'title': 'Item:Q10', 'lastr
evid': 4682, 'modified': '2020-05-08T15:25:16Z', 'type': 'item', 'id': 'Q1
0', 'labels': {'fr': {'language': 'fr', 'value': 'Andrée de Jongh'}}, 'des
criptions': {'fr': {'language': 'fr', 'value': 'résistante belge'}}, 'alia
ses': {'fr': [{'language': 'fr', 'value': 'DEDE'}, {'language': 'fr', 'val
ue': 'Countess Andrée de Jongh'}, {'language': 'fr', 'value': 'the Postma
n'}, {'language': 'fr', 'value': 'Dédée'}, {'language': 'fr', 'value': 'Co
mtesse Andrée de Jongh'}]}, 'claims': {'P2': [{'mainsnak': {'snaktype': 'v
alue', 'property': 'P2', 'hash': '1e39a37398e29ca8e51667fd72bf5035feaa62f
7', 'datavalue': {'value': 'Q461027', 'type': 'string'}, 'datatype': 'exte
rnal-id'}, 'type': 'statement', 'id': 'Q10$86e7777d-48e3-0e41-7b43-9363dcb
fd7ad', 'rank': 'normal'}], 'P47': [{'mainsnak': {'snaktype': 'value', 'pr
operty': 'P47', 'hash': '8366b12552e04447f2788e8b886f3eb456c8a3f9', 'datav
alue': {'value': '0000 0000 3783 5977', 'type': 'string'}, 'datatype': 'ex
ternal-id'}, 'type': 'statement', 'id': 'Q10$1c49accd-4579-fc25-6a7f-bf6c9
8e2e28e', 'rank': 'normal'}], 'P48': [{'mainsnak': {'snaktype': 'value',
'property': 'P48', 'hash': 'c0516fca8d78a5c2c7961e1b8b36947f70df2c9e', 'da
tavalue': {'value': '23987033', 'type': 'string'}, 'datatype': 'external-i
d'}, 'type': 'statement', 'id': 'Q10$74a0b869-46a8-a95e-07c3-12e4fc56c3e
9', 'rank': 'normal'}], 'P31': [{'mainsnak': {'snaktype': 'value', 'proper
ty': 'P31', 'hash': 'b64b590f4b457f43b045c5384ef1b05c5a5ad192', 'datavalu
e': {'value': {'time': '+1916-11-30T00:00:00Z', 'timezone': 0, 'before':
0, 'after': 0, 'precision': 11, 'calendarmodel': 'http://www.wikidata.org/
entity/Q1985727'}, 'type': 'time'}, 'datatype': 'time'}, 'type': 'statemen
t', 'id': 'Q10$2492e61e-4038-619f-836b-19a202e41b0b', 'rank': 'normal', 'r
eferences': [{'hash': '856c433babfc78a50119682546733454e55d0c88', 'snaks':
{'P60': [{'snaktype': 'value', 'property': 'P60', 'hash': '03c71c13fae0811
30f53160cad41493c93ecb5ee', 'datavalue': {'value': 'Q461027', 'type': 'str
ing'}, 'datatype': 'external-id'}]}, 'snaks-order': ['P60']}]}], 'P1':
[{'mainsnak': {'snaktype': 'value', 'property': 'P1', 'hash': 'c516a1ecd47
48187475cb6f606b4301eb92106d4', 'datavalue': {'value': {'entity-type': 'it
em', 'numeric-id': 3617, 'id': 'Q3617'}, 'type': 'wikibase-entityid'}, 'da
tatype': 'wikibase-item'}, 'type': 'statement', 'id': 'Q10$5404acfa-4b9a-f
5f9-1799-a5ae583d42bd', 'rank': 'normal'}]}, 'sitelinks': {}}}, 'success':
1}
```

Dataframe Pandas

Nous allons profiter de la librairie *Pandas* pour tenter de visualiser ces données JSON de façon peut-être plus lisible... Mais il faut se rappeler que ces données sont avant tout destinées à être lisibles par des machines : la valeur des propriétés (colonne 'claims') n'apparaît d'ailleurs pas directement dans le dataframe ci-dessous.

In [4]:

```
DeJongh = pd.DataFrame(result1['entities']["Q10"])
DeJongh.head() # pour voir tout l'entièreté du dataframe, supprimer le _.head()_
```

Out[4]:

	pageid	ns	title	lastrevid	modified	type	id	labels	descriptions	
fr	13	120	Item:Q10	4682	2020-05-08T15:25:16Z	item	Q10	{'language': 'fr', 'value': 'Andrée de Jongh'}	{'language': 'fr', 'value': 'résistante belge'}	[{'la 'fr {'
P2	13	120	Item:Q10	4682	2020-05-08T15:25:16Z	item	Q10	NaN	NaN	
P47	13	120	Item:Q10	4682	2020-05-08T15:25:16Z	item	Q10	NaN	NaN	
P48	13	120	Item:Q10	4682	2020-05-08T15:25:16Z	item	Q10	NaN	NaN	
P31	13	120	Item:Q10	4682	2020-05-08T15:25:16Z	item	Q10	NaN	NaN	

Valeur d'une propriété

Pour obtenir la valeur d'une certaine propriété (colonne *claims*), nous pouvons donc revenir au JSON en ciblant ce que l'on souhaite afficher (ici une **date de naissance** (P31) qui apparaît après *datavalue*) :

In [5]:

```
result1['entities']['Q10']['claims']['P31']
```

Out[5]:

```
[{'mainsnak': {'snaktype': 'value',
   'property': 'P31',
   'hash': 'b64b590f4b457f43b045c5384ef1b05c5a5ad192',
   'datavalue': {'value': {'time': '+1916-11-30T00:00:00Z',
     'timezone': 0,
     'before': 0,
     'after': 0,
     'precision': 11,
     'calendarmodel': 'http://www.wikidata.org/entity/Q1985727'},
    'type': 'time'},
   'datatype': 'time'},
  'type': 'statement',
  'id': 'Q10$2492e61e-4038-619f-836b-19a202e41b0b',
  'rank': 'normal',
  'references': [{'hash': '856c433babfc78a50119682546733454e55d0c88',
    'snaks': {'P60': [{'snaktype': 'value',
       'property': 'P60',
       'hash': '03c71c13fae081130f53160cad41493c93ecb5ee',
       'datavalue': {'value': 'Q461027', 'type': 'string'},
       'datatype': 'external-id'}]},
    'snaks-order': ['P60']}]}]
```

Parsing avancé

Comme vous pouvez le constater ci-dessus, l'extraction de valeurs associées à une propriété en particulier se révèle complexe en raison de la structure du JSON utilisée par Wikidata (voir par exemple ce post de Steve Baskauf) (http://baskauf.blogspot.com/2019/06/putting-data-into-wikidata-using.html).

En raison des spécificités du modèle de données Wikidata et de la complexité induite par les différents types de données (data-type) (https://www.wikidata.org/wiki/Help:Data_type/fr) associées aux propriétés Wikidata - qui nécessiteraient des traitements sur mesure -, nous n'approfondissons donc pas ici le *parsing* de données récursives et conseillons plutôt de se référer à des outils pré-existants comme wikibase-cli (the Command-line interface interface to Wikibase instances) (https://github.com/maxlath/wikibase-cli), Wikidata Integrator exemple (https://colab.research.google.com/drive/1MetrPsLghOmrD-igw7bQlibZ3IdiDVya) ou Wikidata GraphQL voir cette version de Tpt (https://tools.wmflabs.org/tptools/) + exemples (https://phabricator.wikimedia.org/T173214) ou celle-ci (avec exemples) (https://github.com/lisongx/wikidata-graphql). Il pourrait également être envisagé de passer par OpenRefine (https://groups.google.com/forum/#!searchin/openrefine/jython|sort:date/openrefine/x7llXdOTHu8/rNONaZuFA/

Extraire les données de plusieurs entités

Il est également possible de rechercher de lancer un appel à l'API pour plusieurs entités, c'est également possible.

Admettons par exemple que l'on recherche les descriptions de différentes communes belges (Q100, Q101, Q103, Q104) dans toutes les langues disponibles :

In [6]:

```
get_communes = 'https://adochs.arch.be/w/api.php?action=wbgetentities&ids=Q100|Q101|Q10
3|Q104&props=descriptions&format=json'
res = requests.get(get_communes, verify=SSL_VERIFY)
result2 = res.json()
```

In [7]:

```
print(result2)
```

```
{'entities': {'Q100': {'type': 'item', 'id': 'Q100', 'descriptions': {'f
r': {'language': 'fr', 'value': "commune de la province d'Anvers (Belgiqu
e)"}, 'nl': {'language': 'nl', 'value': 'gemeente van de provincie Antwerp
en (België)'}, 'de': {'language': 'de', 'value': 'Gemeinde der Provinz Ant
werpen (Belgien)'}, 'en': {'language': 'en', 'value': 'municipality of the
province of Antwerp (Belgium)'}}}, 'Q101': {'type': 'item', 'id': 'Q101',
'descriptions': {'fr': {'language': 'fr', 'value': "commune de la province
d'Anvers (Belgique)"}, 'nl': {'language': 'nl', 'value': 'gemeente van de
provincie Antwerpen (België)'}, 'de': {'language': 'de', 'value': 'Gemeind
e der Provinz Antwerpen (Belgien)'}, 'en': {'language': 'en', 'value': 'mu
nicipality of the province of Antwerp (Belgium)'}}}, 'Q103': {'type': 'ite
m', 'id': 'Q103', 'descriptions': {'fr': {'language': 'fr', 'value': "comm
une de la province d'Anvers (Belgique)"}, 'nl': {'language': 'nl', 'valu
e': 'gemeente van de provincie Antwerpen (België)'}, 'de': {'language': 'd
e', 'value': 'Gemeinde der Provinz Antwerpen (Belgien)'}, 'en': {'languag
e': 'en', 'value': 'municipality of the province of Antwerp (Belgium)'}}},
'Q104': {'type': 'item', 'id': 'Q104', 'descriptions': {'fr': {'language':
'fr', 'value': "commune de la province d'Anvers (Belgique)"}, 'nl': {'lang
uage': 'nl', 'value': 'gemeente van de provincie Antwerpen (België)'}, 'd
e': {'language': 'de', 'value': 'Gemeinde der Provinz Antwerpen (Belgie
n)'}, 'en': {'language': 'en', 'value': 'municipality of the province of A
ntwerp (Belgium)'}}}}, 'success': 1}
```

1.2 Rechercher une entité

Dans cet exemple, nous lançons un appel à l'API à l'aide de l'action **wbsearchentities** afin de voir si la Wikibase contient une entité associée à *the Postman*, qui n'est autre que l'un des noms de code qui fut porté par Andrée de Jongh dans le cadre des ses activités de résistance.

Comme ce nom de code a été documenté parmi ses alias (https://adochs.arch.be/w/index.php?title=Item:Q10), l'entité Q10 de la Wikibase devrait normalement nous être renvoyée.

NB : il est obligatoire de préciser la langue de recherche et l'API ne semble pas utiliser de *fuzzy matching*, il faut donc que la même graphie soit utilisée que celle encodée dans la Wikibase.

In [8]:

```
find_postman = 'https://adochs.arch.be/w/api.php?action=wbsearchentities&format=json&search=the Postman&language=fr'
res = requests.get(find_postman, verify=SSL_VERIFY)
result3 = res.json()
```

In [9]:

```
result3
```

Out[9]:

```
{'searchinfo': {'search': 'the Postman'},
 'search': [{'repository': '',
   'id': 'Q10',
   'concepturi': 'https://adochs.arch.be/entity/Q10',
   'title': 'Item:Q10',
   'pageid': 13,
   'url': 'https://adochs.arch.be/wiki/Item:Q10',
   'match': {'type': 'alias', 'language': 'fr', 'text': 'the Postman'},
   'aliases': ['the Postman']}],
 'success': 1}
```

In [10]:

```
ThePostman = pd.DataFrame(result3["search"])
ThePostman.head()
```

Out[10]:

	repository	id	concepturi	title	pageid	
0		Q10	https://adochs.arch.be/entity/Q10	Item:Q10	13	https://adochs.arch.be/wiki/I

2. SPARQL Endpoint

Ce deuxième point est directement inspiré d'un script développé par Steve Baskauf (the Vanderbilt Libraries) ainsi que de ses posts de blogs (Getting Data Out of Wikidata using Software (http://baskauf.blogspot.com/2019/05/getting-data-out-of-wikidata-using.html) et, dans une moindre mesure SPARQL: Retrieving SPARQL query data using HTTP (https://heardlibrary.github.io/digital-scholarship/lod/sparql/#retrieving-sparql-query-data-using-http)). L'idée est d'utiliser le service de requêtes de la Wikibase comme une API en tirant parti d'un script Python.

Ici nous utiliserons plus spécifiquement la librairie Pandas (https://pandas.pydata.org/), *a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.*

Le code ci-dessous est inspiré de :

- Building a stand-alone off-Wiki layered map using Wikidata & SPARQL (https://paws-public.wmflabs.org/paws-public/10180704/WikidataMapMakingWorkshop.ipynb)
- Working with pywikibot, Wikidata API, and SPARQL (https://github.com/remerjohnson/wikidata-analysis/blob/master/wikidata_pywikibot_sparql_work.ipynb)
- SPARQL to Pandas Dataframes (https://lawlesst.github.io/notebook/sparql-dataframe.html)
- Querying a SPARQL endpoint in Python (https://github.com/SuLab/sparql_to_pandas/blob/master/SPARQL_pandas.ipynb)
- Sunday Query : use SPARQL and Python to fix typographical errors on Wikidata (https://blog.ash.bzh/en/sunday-query-use-sparql-and-python-to-fix-typographical-errors-on-wikidata)

Prérequis : installer SPARQLWrapper (https://sparqlwrapper.readthedocs.io/en/latest/main.html), peut être fait via Anaconda :

```
coda install -c conda-forge sparqlwrapper
```

In [11]:

```
import pandas as pd
import json
from SPARQLWrapper import SPARQLWrapper, JSON
```

Objectif

Le but est d'extraire la liste de toutes les communes belges stockées dans la Wikibase, avec leur libellé et leur description en français, ainsi que leur identifiant AGR et leur code INS.

In [12]:

```
endpoint_url = "https://query-adochs.arch.be/proxy/wdqs/bigdata/namespace/wdq/sparql"
#si nous interrogions Wikidata, ça serait "https://query.wikidata.org/sparql"
```

In [13]:

```
#Le code d'une requête peut être obtenue sur l'interface graphique du SPARQL endpoint e
n cliquant en bas à droite sur </>code
# et en choisissant ensuite 'Python', voir l'exemple ci-dessous: https://tinyurl.com/yb
22fbym

query = """
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>

SELECT ?place ?placeLabel ?placeDescription ?identifiantAGR ?codeINS ?identifiantWikida
ta WHERE {
    ?place wbt:P1 wb:Q25. #je cherche des communes
    ?place wbt:P57 ?identifiantAGR.
  OPTIONAL { ?place wbt:P53 ?codeINS. }
  OPTIONAL { ?place wbt:P2 ?identifiantWikidata. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,nl" } .
}
ORDER BY xsd:integer (?identifiantAGR) """
```

Nous présentons en détail le processus pour récupérer les données au **format JSON sous le point A/**.

Nous présentons également une méthode (inspirée de cette fonction (https://github.com/lawlesst/lawlesst.github.com/blob/pelican/content/jupyter/sparql_dataframe.ipynb)) pour stocker les données dans un **dataframe Pandas (exportable en fichier CSV) sous le point B**.

A/ Obtenir les données en JSON

Pour obtenir des détails sur la fonction qui suit, se référer à ce post (https://blog.ash.bzh/en/sunday-query-use-sparql-and-python-to-fix-typographical-errors-on-wikidata/).

In [14]:

```
def get_results(endpoint_url, query):
    sparql = SPARQLWrapper(endpoint_url, agent='User:Annette')
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()

results = get_results(endpoint_url, query)

#print(results)
```

Ci-dessous, un petit extrait de ce qu'on obtient lorsqu'on lance la commande *print(results) :

{'head': {'vars': ['place', 'placeLabel', 'placeDescription', 'identifiantAGR', 'codeINS', 'identifiantWikidata']}, 'results': {'bindings': [{'place': {'type': 'uri', 'value': 'https://adochs.arch.be/entity/Q39'}, 'identifiantAGR': {'type': 'literal', 'value': '2'}, 'codeINS': {'type': 'literal', 'value': '11002'}, 'identifiantWikidata': {'type': 'literal', 'value': 'Q12892'}, 'placeLabel': {'xml:lang': 'fr', 'type': 'literal', 'value': 'Anvers'}, 'placeDescription': {'xml:lang': 'fr', 'type': 'literal', 'value': "commune de la province d'Anvers (Belgique)"}}, {'place': {'type': 'uri', 'value': 'https://adochs.arch.be/entity/Q67'}, 'identifiantAGR': {'type': 'literal', 'value': '19'}, 'codeINS': {'type': 'literal', 'value':

'11004'}, 'identifiantWikidata': {'type': 'literal', 'value': 'Q723822'}, 'placeLabel': {'xml:lang': 'fr', 'type': 'literal', 'value': 'Boechout'}, 'placeDescription': {'xml:lang': 'fr', 'type': 'literal', 'value': "commune de la province d'Anvers (Belgique)"}},

Si nous regardons le contenu de cet extrait, nous voyons qu'il s'agit d'un **dictionnaire Python avec deux entrées** : *head* et *results*.

- L'entrée ***head*** contient le noms des six variables renvoyées par la requête (place, placeLabel, etc.).
- L'entrée ***results*** contient un autre dictionnaire, contenant la clé *bindings*, contenant une liste des résultats en tant que tels, présents chacun dans un nouveau dictionnaire Python.

Si nous observons les premiers résultats correspondant à l'entité Anvers|Q39 (https://adochs.arch.be/entity/Q39), nous pouvons inventorier ce que ce dictionnaire contient :

- les 6 clés correspondants aux variables renvoyées par la requête *(place, identifiantAGR, codeINS, identifiantWikidata, placeDescription)*
- **pour chacune de ces 6 clés**, une valeur correspondant à un autre dictionnaire (...), dont **la clé *value* contient la valeur que nous recherchons** !

L'étape suivante vise donc à *parser* les résultats afin de ne garder que ce qui nous intéresse.

Parser les résultats

Cette étape vise à ***parser* la liste *bindings*** à l'aide d'une boucle Python (*for*) afin d'extraire la valeur recherchée. Nous allons extraire de chaque dictionnaire de résultat *(results["results"]["bindings"])* la valeur de *value*.

In [15]:

```
for result in results["results"]["bindings"]:
    place = result['place']['value']
    label = result['placeLabel']['value']
    description = result['placeDescription']['value']
    AGR = result['identifiantAGR']['value']
    INS = result['codeINS']['value']
    Wikidata = result['identifiantWikidata']['value']

    print(place, label, description, AGR, INS, Wikidata)
```

https://adochs.arch.be/entity/Q39 Anvers commune de la province d'Anvers (Belgique) 2 11002 Q12892
https://adochs.arch.be/entity/Q67 Boechout commune de la province d'Anvers (Belgique) 19 11004 Q723822
https://adochs.arch.be/entity/Q66 Boom commune de la province d'Anvers (Belgique) 22 11005 Q723972
https://adochs.arch.be/entity/Q68 Borsbeek commune de la province d'Anvers (Belgique) 23 11007 Q724055
https://adochs.arch.be/entity/Q85 Brasschaat commune de la province d'Anvers (Belgique) 24 11008 Q693513
https://adochs.arch.be/entity/Q83 Brecht commune de la province d'Anvers (Belgique) 25 11009 Q724131
https://adochs.arch.be/entity/Q75 Edegem commune de la province d'Anvers (Belgique) 29 11013 Q724238
https://adochs.arch.be/entity/Q73 Essen commune de la province d'Anvers (Belgique) 30 11016 Q724283
https://adochs.arch.be/entity/Q121 Hemiksem commune de la province d'Anvers (Belgique) 31 11018 Q724448
https://adochs.arch.be/entity/Q115 Hove commune de la province d'Anvers (Belgique) 32 11021 Q649982
https://adochs.arch.be/entity/Q77 Calmpthout commune de la province d'Anvers (Belgique) 33 11022 Q724777
https://adochs.arch.be/entity/Q84 Kapellen commune de la province d'Anvers (Belgique) 34 11023 Q724942
https://adochs.arch.be/entity/Q69 Kontich commune de la province d'Anvers (Belgique) 37 11024 Q725074
https://adochs.arch.be/entity/Q72 Lint commune de la province d'Anvers (Belgique) 40 11025 Q725268
https://adochs.arch.be/entity/Q74 Mortsel commune de la province d'Anvers (Belgique) 41 11029 Q688781
https://adochs.arch.be/entity/Q76 Niel commune de la province d'Anvers (Belgique) 42 11030 Q912054
https://adochs.arch.be/entity/Q120 Ranst commune de la province d'Anvers (Belgique) 43 11035 Q501748
https://adochs.arch.be/entity/Q114 Rumst commune de la province d'Anvers (Belgique) 48 11037 Q943446
https://adochs.arch.be/entity/Q106 Schelle commune de la province d'Anvers (Belgique) 52 11038 Q911968
https://adochs.arch.be/entity/Q104 Schilde commune de la province d'Anvers (Belgique) 53 11039 Q263177
https://adochs.arch.be/entity/Q109 Schoten commune de la province d'Anvers (Belgique) 56 11040 Q655203
https://adochs.arch.be/entity/Q110 Stabroek commune de la province d'Anvers (Belgique) 57 11044 Q595654
https://adochs.arch.be/entity/Q103 Wijnegem commune de la province d'Anvers (Belgique) 60 11050 Q527808
https://adochs.arch.be/entity/Q101 Wommelgem commune de la province d'Anvers (Belgique) 61 11052 Q839037
https://adochs.arch.be/entity/Q100 Wuustwezel commune de la province d'Anvers (Belgique) 62 11053 Q912045
https://adochs.arch.be/entity/Q97 Zandhoven commune de la province d'Anvers (Belgique) 65 11054 Q146656
https://adochs.arch.be/entity/Q86 Zoersel commune de la province d'Anvers (Belgique) 71 11055 Q218253
https://adochs.arch.be/entity/Q124 Zwijndrecht commune de la province d'Anvers (Belgique) 74 11056 Q244690
https://adochs.arch.be/entity/Q127 Malle commune de la province d'Anvers (Belgique) 77 11057 Q917591

In [16]:

```
#!/usr/bin/env python3

from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query-adochs.arch.be/proxy/wdqs/bigdata/namespace/wdq/sparql"

query = """
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>

SELECT ?place ?placeLabel ?placeDescription ?identifiantAGR ?codeINS ?identifiantWikida
ta WHERE {
    ?place wbt:P1 wb:Q25. #je cherche des communes
    ?place wbt:P57 ?identifiantAGR.
  OPTIONAL { ?place wbt:P53 ?codeINS. }
  OPTIONAL { ?place wbt:P2 ?identifiantWikidata. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,nl" } .
}
ORDER BY xsd:integer (?identifiantAGR) """

def get_results(endpoint_url, query):
    sparql = SPARQLWrapper(endpoint_url)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()

results = get_results(endpoint_url, query)


for result in results["results"]["bindings"]:
    place = result['place']['value']
    label = result['placeLabel']['value']
    description = result['placeDescription']['value']
    AGR = result['identifiantAGR']['value']
    INS = result['codeINS']['value']
    Wikidata = result['identifiantWikidata']['value']

    print(place, label, description, AGR, INS, Wikidata)
```

https://adochs.arch.be/entity/Q39 Anvers commune de la province d'Anvers (Belgique) 2 11002 Q12892
https://adochs.arch.be/entity/Q67 Boechout commune de la province d'Anvers (Belgique) 19 11004 Q723822
https://adochs.arch.be/entity/Q66 Boom commune de la province d'Anvers (Belgique) 22 11005 Q723972
https://adochs.arch.be/entity/Q68 Borsbeek commune de la province d'Anvers (Belgique) 23 11007 Q724055
https://adochs.arch.be/entity/Q85 Brasschaat commune de la province d'Anvers (Belgique) 24 11008 Q693513
https://adochs.arch.be/entity/Q83 Brecht commune de la province d'Anvers (Belgique) 25 11009 Q724131
https://adochs.arch.be/entity/Q75 Edegem commune de la province d'Anvers (Belgique) 29 11013 Q724238
https://adochs.arch.be/entity/Q73 Essen commune de la province d'Anvers (Belgique) 30 11016 Q724283
https://adochs.arch.be/entity/Q121 Hemiksem commune de la province d'Anvers (Belgique) 31 11018 Q724448
https://adochs.arch.be/entity/Q115 Hove commune de la province d'Anvers (Belgique) 32 11021 Q649982
https://adochs.arch.be/entity/Q77 Calmpthout commune de la province d'Anvers (Belgique) 33 11022 Q724777
https://adochs.arch.be/entity/Q84 Kapellen commune de la province d'Anvers (Belgique) 34 11023 Q724942
https://adochs.arch.be/entity/Q69 Kontich commune de la province d'Anvers (Belgique) 37 11024 Q725074
https://adochs.arch.be/entity/Q72 Lint commune de la province d'Anvers (Belgique) 40 11025 Q725268
https://adochs.arch.be/entity/Q74 Mortsel commune de la province d'Anvers (Belgique) 41 11029 Q688781
https://adochs.arch.be/entity/Q76 Niel commune de la province d'Anvers (Belgique) 42 11030 Q912054
https://adochs.arch.be/entity/Q120 Ranst commune de la province d'Anvers (Belgique) 43 11035 Q501748
https://adochs.arch.be/entity/Q114 Rumst commune de la province d'Anvers (Belgique) 48 11037 Q943446
https://adochs.arch.be/entity/Q106 Schelle commune de la province d'Anvers (Belgique) 52 11038 Q911968
https://adochs.arch.be/entity/Q104 Schilde commune de la province d'Anvers (Belgique) 53 11039 Q263177
https://adochs.arch.be/entity/Q109 Schoten commune de la province d'Anvers (Belgique) 56 11040 Q655203
https://adochs.arch.be/entity/Q110 Stabroek commune de la province d'Anvers (Belgique) 57 11044 Q595654
https://adochs.arch.be/entity/Q103 Wijnegem commune de la province d'Anvers (Belgique) 60 11050 Q527808
https://adochs.arch.be/entity/Q101 Wommelgem commune de la province d'Anvers (Belgique) 61 11052 Q839037
https://adochs.arch.be/entity/Q100 Wuustwezel commune de la province d'Anvers (Belgique) 62 11053 Q912045
https://adochs.arch.be/entity/Q97 Zandhoven commune de la province d'Anvers (Belgique) 65 11054 Q146656
https://adochs.arch.be/entity/Q86 Zoersel commune de la province d'Anvers (Belgique) 71 11055 Q218253
https://adochs.arch.be/entity/Q124 Zwijndrecht commune de la province d'Anvers (Belgique) 74 11056 Q244690
https://adochs.arch.be/entity/Q127 Malle commune de la province d'Anvers (Belgique) 77 11057 Q917591

```
a province du Brabant wallon (Belgique) 651 25121 Q329642
https://adochs.arch.be/entity/Q404 Ramillies commune de la province du Bra
bant wallon (Belgique) 657 25122 Q531456
https://adochs.arch.be/entity/Q405 Rebecq commune de la province du Braban
t wallon (Belgique) 669 25123 Q374861
https://adochs.arch.be/entity/Q406 Walhain commune de la province du Braba
nt wallon (Belgique) 673 25124 Q650216
https://adochs.arch.be/entity/Q407 Beernem commune de la province de Fland
re occidentale (Belgique) 679 31003 Q687921
https://adochs.arch.be/entity/Q408 Blankenberge commune de la province de
Flandre occidentale (Belgique) 683 31004 Q328104

---------------------------------------------------------------------------
-
KeyError                                  Traceback (most recent call las
t)
<ipython-input-16-3ef23b6a5cf8> in <module>
     33     AGR = result['identifiantAGR']['value']
     34     INS = result['codeINS']['value']
---> 35     Wikidata = result['identifiantWikidata']['value']
     36
     37     print(place, label, description, AGR, INS, Wikidata)

KeyError: 'identifiantWikidata'
```

B) Charger les résultats dans un dataframe Pandas

NB : Pour éviter les redondances, nous réutilisons les variables (*endpoint_url* et *query*, voir ci-dessus).

In [17]:

```
def get_sparql_dataframe(endpoint_url, query):
    """
    Helper function to convert SPARQL results into a Pandas data frame.
    """
    sparql = SPARQLWrapper(endpoint_url)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    result = sparql.query()

    processed_results = json.load(result.response)
    cols = processed_results['head']['vars']

    out = []
    for row in processed_results['results']['bindings']:
        item = []
        for c in cols:
            item.append(row.get(c, {}).get('value'))
        out.append(item)

    return pd.DataFrame(out, columns=cols)
```

In [18]:

```
df_places = get_sparql_dataframe(endpoint_url, query)
df_places.head()
```

Out[18]:

	place	placeLabel	placeDescription	identifiantAGR	codeINS	ident
0	https://adochs.arch.be/entity/Q39	Anvers	commune de la province d'Anvers (Belgique)	2	11002	
1	https://adochs.arch.be/entity/Q67	Boechout	commune de la province d'Anvers (Belgique)	19	11004	
2	https://adochs.arch.be/entity/Q66	Boom	commune de la province d'Anvers (Belgique)	22	11005	
3	https://adochs.arch.be/entity/Q68	Borsbeek	commune de la province d'Anvers (Belgique)	23	11007	
4	https://adochs.arch.be/entity/Q85	Brasschaat	commune de la province d'Anvers (Belgique)	24	11008	

Exporter les données dans un fichier CSV

In [19]:

```
df_places.to_csv("df_places.csv", sep = '\t', index = False, encoding = 'utf-8')
```

Script final B/

(Version condensée des étapes précédentes)

In [20]:

```
#!/usr/bin/env python3

import pandas as pd
from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query-adochs.arch.be/proxy/wdqs/bigdata/namespace/wdq/sparql"

query = """
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>

SELECT ?place ?placeLabel ?placeDescription ?identifiantAGR ?codeINS ?identifiantWikida
ta WHERE {
    ?place wbt:P1 wb:Q25. #je cherche des communes
    ?place wbt:P57 ?identifiantAGR.
  OPTIONAL { ?place wbt:P53 ?codeINS. }
  OPTIONAL { ?place wbt:P2 ?identifiantWikidata. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,nl" } .
}
ORDER BY xsd:integer (?identifiantAGR) """

def get_sparql_dataframe(endpoint_url, query):
    """
    Helper function to convert SPARQL results into a Pandas data frame.
    """
    sparql = SPARQLWrapper(endpoint_url)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    result = sparql.query()

    processed_results = json.load(result.response)
    cols = processed_results['head']['vars']

    out = []
    for row in processed_results['results']['bindings']:
        item = []
        for c in cols:
            item.append(row.get(c, {}).get('value'))
        out.append(item)

    return pd.DataFrame(out, columns=cols)

df_places = get_sparql_dataframe(endpoint_url, query)
```

Aller plus loin

Évidemment, il est possible de lancer des requêtes plus complexes, croisant davantage d'éléments et de propriétés, par exemple dans le but de générer une *heat map* basée sur des données contenues dans la Wikibase ici (https://paws-public.wmflabs.org/paws-public/User:OlafJanssen/WikidataMapMakingWorkshop/WikidataMapMakingWorkshop.ipynb) ou encore visant à récupérer un identifiant Wikidata à partir d'un identifiant VIAF ici (https://github.com/remerjohnson/wikidata-analysis/blob/master/wikidata_pywikibot_sparql_work.ipynb).

De plus, une fois que les données sont chargées dans un dataframe Pandas, il est possible d'en profiter pour faire d'autres choses, comme par exemple compter le nombre de personnes partageant la même description ici (https://github.com/lawlesst/lawlesst.github.com/blob/pelican/content/jupyter/sparql_dataframe.ipynb), ré-ordonner l'ordre des éléments, combiner des colonnes, etc. ici (https://github.com/SuLab/sparql_to_pandas/blob/master/SPARQL_pandas.ipynb), .

Annexe 10
Code [157] Python permettant la conversion d'éléments Wikibase en notices EAC-CPF

[ ]:
```
# NOTE: run in python3
# Script créé par Adrien di Mascio, modifié par Anne Chardonnens.


from flask import Flask, Response
from SPARQLWrapper import SPARQLWrapper, JSON


def query_wikibase(qid):

    sparql = SPARQLWrapper("https://query-adochs.arch.be/proxy/wdqs/bigdata/
 ↪namespace/wdq/sparql") #("https://query.wikidata.org/sparql")
    sparql.setQuery("""


PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>

select distinct ?personneLabel ?personneDescription ?nomLabel ?prenomLabel ?
 ↪date_naissanceLabel ?date_mortLabel ?lieu_naissanceLabel ?lieu_mortLabel ?
 ↪genreLabel ?nationaliteLabel ?metierLabel ?viaf ?isni
where {

BIND (wb:%s as ?personne) .

  OPTIONAL  {?personne wdt:P67 ?nom.}
  ?personne wbt:P31 ?date_naissance.

   OPTIONAL {?personne wbt:P32 ?date_mort.}
   OPTIONAL {?personne wbt:P66 ?prenom.}
   OPTIONAL {?personne wbt:P33 ?lieu_naissance.}
   OPTIONAL {?personne wbt:P34 ?lieu_mort.}
   OPTIONAL {?personne wbt:P37 ?genre.}
   OPTIONAL {?personne wbt:P38 ?nationalite.}
   OPTIONAL {?personne wbt:P54 ?metier.}
   OPTIONAL {?personne wbt:P48 ?viaf.}
   OPTIONAL {?personne wbt:P47 ?isni.}
```

157. Adaptation d'un script initialement créé par Adrien di Mascio dans le cadre du premier hackathon des Archives nationales de France.

```
  service wikibase:label { bd:serviceParam wikibase:language␣
↪"[AUTO_LANGUAGE],fr,en,nl". }
}

        """ % qid)

    sparql.setReturnFormat(JSON)
    results = sparql.query().convert()
    processed = {}

    for var in results["head"]["vars"]:
        try:
            processed[var] = results["results"]["bindings"][0][var]["value"]
        except (IndexError, KeyError) as err:
            processed[var] = ""
    return processed


app = Flask(__name__)
template = open('eac_template_arch.xml').read()



BASE_INFOS = {
    'prenom_artiste': '',
    'annee_naissance': '',
    'date_naissance': '',
    'date_mort': '',
    'annee_mort': '',
    'annee_commande': '',
    'lieu_naissance': '',
    'lieu_mort': '',
    'genreLabel': '',
   'nationalite': '',
    'metier': '',
    'q': '',
    'bnf': '',
    'viaf': '',
    'isni': '',
    'odis': '',
    'rkd': '',
    'ulan': '',
    'snac': '',
```

```
    'URL_VIAF': '',
    'ark_bnf': '',

}


@app.route('/eac/<qid>')
def generate_eac(qid):
    infos = BASE_INFOS.copy()
    infos['q'] = 'https://adochs.arch.be/entity/%s' % qid #'http://www.wikidata.
 ↪org/entity/%s' % qid
    infos.update(query_wikibase(qid))
    if infos['date_mortLabel']:
        infos['annee_mort'] = infos['date_mortLabel'][:4]
    if infos['date_naissanceLabel']:
        infos['annee_naissance'] = infos['date_naissanceLabel'][:4]
    if infos['date_mortLabel']:
        infos['date_mortLabel'] = infos['date_mortLabel'][:10]
    if infos['date_naissanceLabel']:
        infos['date_naissanceLabel'] = infos['date_naissanceLabel'][:10]
    if not infos['nomLabel'] and infos['personneLabel']:
        infos['prenomLabel'] = ''
        infos['nomLabel'] = infos['personneLabel']
    data = template % infos
    return Response(data, mimetype='text/xml')


@app.route('/')
def home():
    return 'hello'
```

Annexe 11
Requêtes SPARQL fédérées

Pour plus d'infos, consulter cette page (https://linkingthepast.org/about/).

Effectuer des requêtes SPARQL fédérées

Cette étape vise à explorer comment nous pouvons interroger simultanément plusieurs points d'accès SPARQL.

Objectif

Le but est de tester l'utilisation de requêtes SPARQL fédérées, pour croiser les données de la Wikibase avec des données externes. En l'occurence, le test présenté ci-dessous vise à compléter les données sur une personne à l'aide de donnée issues de Wikidata.

Inspiration

Les exemples ci-dessous sont inspirés de ces deux requêtes fédérées :

- Biodiversity + Wikidata (https://tinyurl.com/ydawxohx)
- Wikidata + Nobelprize (https://w.wiki/VW7)

Fonctionnement

L'idée est d'utiliser le service de requête (https://query-adochs.arch.be/) de la Wikibase comme pour une requête classique, puis d'utiliser une jointure, c'est-à-dire un pivot permettant de faire le lien entre les données de la Wikibase et les données de Wikidata (dans la cas présent, il s'agit de l'identifiant Wikidata stocké dans la Wikibase à l'aide de propriété P2). La récupération des données complémentaires se fait à l'aide de la "clause" SERVICE, qui permet d'interroger simultanément un autre point d'accès SPARQL.

Nota bene

Ce document vise à présenter le fonctionnement général de telles requêtes, les exemples sont donc illustrés à l'aide du code de la requête et de résultats présentés sous forme de captures d'écran pour une meilleure lisibilité, mais il est bien sûr également possible de lancer ces requêtes à l'aide d'un script Python basé sur le module SPARQLWrapper, à l'instar de ce que nous avons présenté ici (https://linkingthepast.org/reuse/).

Exemple 1

Le but ici est de récupérer une liste des personnes stockées dans la Wikibase *(wbt:P1 wb:Q3617)*, accompagnées des institutions possédant des archives à leur sujet (wdt:P485) (https://www.wikidata.org/wiki/Property:P485) selon Wikidata, ainsi que du nom *(label)* de ces institutions si l'information est disponible.

L'URI de l'identifiant Wikidata *(wbt:P2)* est utilisée comme jointure pour récupérer l'information désirée, tandis que la partie concernée de la requête est transmise au point d'accès de Wikidata à l'aide de la clause SERVICE et de l'URL concernée (https://query.wikidata.org/bigdata/namespace/wdq/).

Les informations sur les institutions possédant des archives au sujet de cette personne (URI Wikidata et nom de l'institution) sont regroupées à l'aide de GROUP_CONTACT pour éviter d'avoir une multiplication de lignes pour une même personne en cas de valeurs multiples.

Requête SPARQL Exemple 1

Rappel: le code ci-dessous est montré à titre purement illustratif mais ne peut pas être lancé dans le cadre de ce Jupyter Notebook.

Pour lancer la requête, suivre ce lien (https://tinyurl.com/ybhcfxds).

In [ ]:

```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wdq: <https://query.wikidata.org/bigdata/namespace/wdq/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?personne ?personneLabel ?wikidata_iri (GROUP_CONCAT(DISTINCT ?archives; SEPA
RATOR = "; ") AS ?archives_URI)
 (GROUP_CONCAT(DISTINCT ?archivesLabel; SEPARATOR = "; ") AS ?archives_label)

WHERE {
  ?personne wbt:P1 wb:Q3617.
  ?personne wbt:P2 ?identifiantWikidata.
  BIND (URI(CONCAT("http://www.wikidata.org/entity/",?identifiantWikidata)) AS ?wikid
ata_iri)
 SERVICE wdq:sparql {
  ?wikidata_iri wdt:P485 ?archives.
  OPTIONAL {?archives rdfs:label ?archivesLabel.
  FILTER(LANG(?archivesLabel) = "[AUTO_LANGUAGE]").}
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,nl,en,de" } .
}

GROUP BY ?personne ?personneLabel ?wikidata_iri ?archives_URI ?archives_label
```

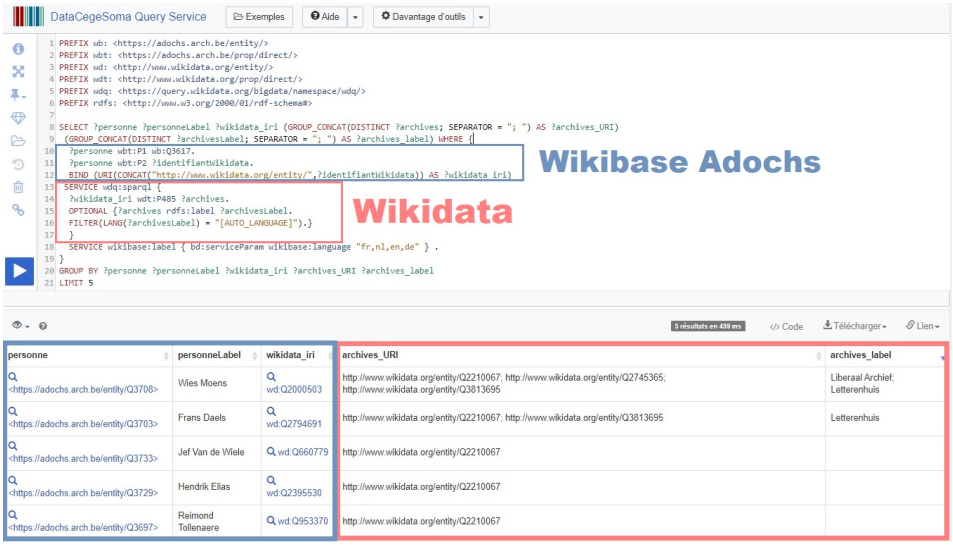
Résultats Exemple 1

personne	personneLabel	wikidata_iri	archives_URI	archives_label
<https://adochs.arch.be/entity/Q3708>	Wies Moens	wd:Q2000503	http://www.wikidata.org/entity/Q2210067; http://www.wikidata.org/entity/Q2745365; http://www.wikidata.org/entity/Q3813695	Liberaal Archief; Letterenhuis
<https://adochs.arch.be/entity/Q3703>	Frans Daels	wd:Q2794691	http://www.wikidata.org/entity/Q2210067; http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3720>	Georges Marlier	wd:Q22109509	http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3698>	Willy Kessels	wd:Q3569164	http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3740>	Paul Colin	wd:Q3370910	http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3748>	Ward Hermans	wd:Q2711435	http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3700>	Cyriel Verschaeve	wd:Q2229015	http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3721>	Maurice De Wilde	wd:Q2278281	http://www.wikidata.org/entity/Q1846753; http://www.wikidata.org/entity/Q3813695	KADOC; Letterenhuis
<https://adochs.arch.be/entity/Q3707>	Victor Leemans	wd:Q129273	http://www.wikidata.org/entity/Q1846753; http://www.wikidata.org/entity/Q3813695	KADOC; Letterenhuis
<https://adochs.arch.be/entity/Q3689>	Henri de Man	wd:Q666890	http://www.wikidata.org/entity/Q1667757; http://www.wikidata.org/entity/Q3813695	Institut international d'histoire sociale; Letterenhuis
<https://adochs.arch.be/entity/Q3734>	Louise de Landsheere	wd:Q11298630	http://www.wikidata.org/entity/Q1023323	Centre d'études guerre et société
<https://adochs.arch.be/entity/Q10>	Andrée de Jongh	wd:Q461027	http://www.wikidata.org/entity/Q1023323	Centre d'études guerre et société

En bref, les résultats nous donnent à voir :

- l'URI Wikibase de la personne concernée
- son label
- l'URI Wikidata permettant de faire le pivot entre l'instance Wikibase et Wikidata
- les URI Wikidata correspondant aux institutions possédant des archives *au sujet de* cette personne
- les labels de ces institutions s'ils sont disponibles en français, néerlandais, anglais ou allemand

Vue d'ensemble (exemple 1)




```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wdq: <https://query.wikidata.org/bigdata/namespace/wdq/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?personne ?personneLabel ?wikidata_iri (GROUP_CONCAT(DISTINCT ?archives; SEPARATOR = "; ") AS ?archives_URI)
 (GROUP_CONCAT(DISTINCT ?archivesLabel; SEPARATOR = "; ") AS ?archives_label) WHERE {
  ?personne wbt:P1 wb:Q3617.
  ?personne wbt:P2 ?identifiantWikidata.
  BIND (URI(CONCAT("http://www.wikidata.org/entity/",?identifiantWikidata)) AS ?wikidata_iri)
 SERVICE wdq:sparql {
  ?wikidata_iri wdt:P485 ?archives.
  OPTIONAL {?archives rdfs:label ?archivesLabel.
  FILTER(LANG(?archivesLabel) = "[AUTO_LANGUAGE]").}
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,nl,en,de" } .
}
GROUP BY ?personne ?personneLabel ?wikidata_iri ?archives_URI ?archives_label
LIMIT 5
```

personne	personneLabel	wikidata_iri	archives_URI	archives_label
<https://adochs.arch.be/entity/Q3708>	Wies Moens	wd:Q2000503	http://www.wikidata.org/entity/Q2210067; http://www.wikidata.org/entity/Q2745365; http://www.wikidata.org/entity/Q3813695	Liberaal Archief; Letterenhuis
<https://adochs.arch.be/entity/Q3703>	Frans Daels	wd:Q2794691	http://www.wikidata.org/entity/Q2210067; http://www.wikidata.org/entity/Q3813695	Letterenhuis
<https://adochs.arch.be/entity/Q3733>	Jef Van de Wiele	wd:Q660779	http://www.wikidata.org/entity/Q2210067	
<https://adochs.arch.be/entity/Q3729>	Hendrik Elias	wd:Q2395530	http://www.wikidata.org/entity/Q2210067	
<https://adochs.arch.be/entity/Q3697>	Reimond Tollenaere	wd:Q953370	http://www.wikidata.org/entity/Q2210067	

Exemple 2

Le but ici est de récupérer, pour une personne donnée (pour des raisons de performance, cf. Limites ci-dessous), les informations présentes sur Wikidata au sujet de ses éventuelles *occupation*, affiliation à un parti politique, distinction reçue, et des éventuelles institutions possédant des archives à son sujet (cf. Exemple 1).

L'URI de l'identifiant Wikidata *(wbt:P2)* est à nouveau utilisée comme jointure pour récupérer l'information désirée, tandis que la partie concernée de la requête est transmise au point d'accès de Wikidata à l'aide de la clause SERVICE et de l'URL concernée (https://query.wikidata.org/bigdata/namespace/wdq/).

Les valeurs multiples sont regroupées dans une même cellule à l'aide de GROUP_CONTACT.

Requête SPARQL Exemple 2

Rappel : le code ci-dessous est montré à titre purement illustratif mais ne peut pas être lancé dans le cadre de ce Jupyter Notebook.

Pour lancer la requête, suivre ce lien (https://tinyurl.com/y868bhgu).

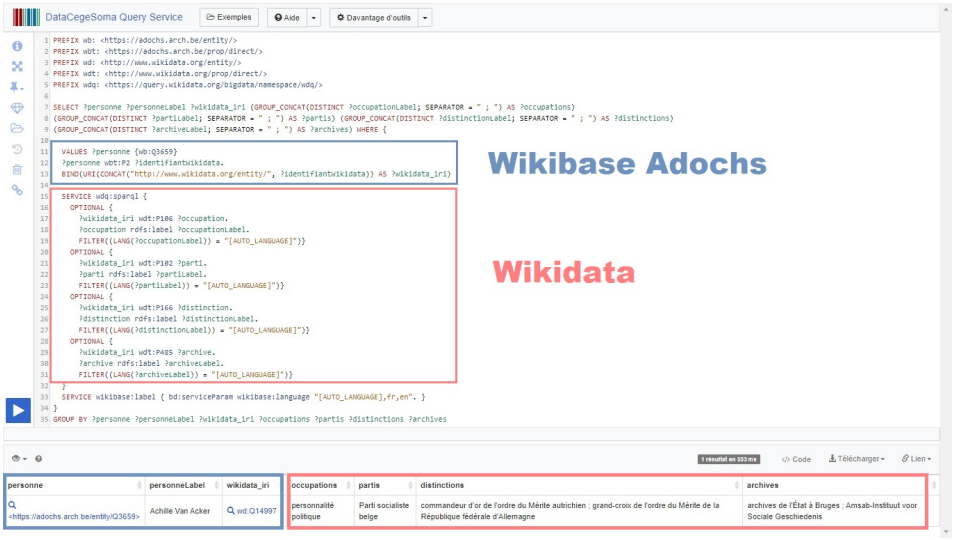
In [ ]:

```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wdq: <https://query.wikidata.org/bigdata/namespace/wdq/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?personne ?personneLabel ?wikidata_iri (GROUP_CONCAT(DISTINCT ?occupationLabel;
SEPARATOR = " ; ") AS ?occupations)
(GROUP_CONCAT(DISTINCT ?partiLabel; SEPARATOR = " ; ") AS ?partis)
(GROUP_CONCAT(DISTINCT ?distinctionLabel; SEPARATOR = " ; ") AS ?distinctions)
(GROUP_CONCAT(DISTINCT ?archiveLabel; SEPARATOR = " ; ") AS ?archives)
WHERE {

  VALUES ?personne {wb:Q3659}
  ?personne wbt:P2 ?identifiantWikidata.

  BIND (URI(CONCAT("http://www.wikidata.org/entity/",?identifiantWikidata)) AS ?wikidat
a_iri)

  SERVICE wdq:sparql {
    OPTIONAL {
      ?wikidata_iri wdt:P106 ?occupation.
      ?occupation rdfs:label ?occupationLabel.
      FILTER((LANG(?occupationLabel)) = "[AUTO_LANGUAGE]")
    }
    OPTIONAL {
      ?wikidata_iri wdt:P102 ?parti.
      ?parti rdfs:label ?partiLabel.
      FILTER((LANG(?partiLabel)) = "[AUTO_LANGUAGE]")
    }
    OPTIONAL {
      ?wikidata_iri wdt:P166 ?distinction.
      ?distinction rdfs:label ?distinctionLabel.
      FILTER((LANG(?distinctionLabel)) = "[AUTO_LANGUAGE]")
    }
       OPTIONAL {
      ?wikidata_iri wdt:P485 ?archive.
      ?archive rdfs:label ?archiveLabel.
      FILTER((LANG(?archiveLabel)) = "[AUTO_LANGUAGE]")
    }
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],fr,en". }
}
GROUP BY ?personne ?personneLabel ?wikidata_iri ?occupations ?partis ?distinctions ?ar
chives
```

Résultats Exemple 2

personne	personneLabel	wikidata_iri	occupations	partis	distinctions	archives
<https://adochs.arch.be/entity/Q3659>	Achille Van Acker	wd:Q14997	personnalité politique	Parti socialiste belge	commandeur d'or de l'ordre du Mérite autrichien ; grand-croix de l'ordre du Mérite de la République fédérale d'Allemagne	archives de l'État à Bruges ; Amsab-Instituut voor Sociale Geschiedenis

Vue d'ensemble (exemple 2)



Limites

Deux principales limites doivent être relevées ici :

- la richesse des résultats est directement conditionnée par l'étendue des données ayant été encodées sur Wikidata (ainsi, une absence de distinctions reçues (https://www.wikidata.org/wiki/Property:P166), ne signifie pas forcément que la personne n'en a reçue aucune...), mais également par la proportion d'entités de la Wikibase ayant pu être réconciliées avec l'identifiant Wikidata correspondant
- lors de nos premiers tests, certains problèmes de performance ont été constatés, nous conduisant à limiter le nombre de résultats à afficher

Aller plus loin

Le processus qui a été illustré ici avec les données issues de Wikidata pourrait également être utilisé pour interroger d'autres points d'accès SPARQL - sous réserve de disposer de données de *jointure* entre ces jeux de données - comme par exemple des points d'accès spécialisés contenant des données liées à la Seconde Guerre mondiale. C'est le par exemple de dati.CDEC (http://dati.cdec.it/indiceEN.html), qui a publié *a database subset of names of Jews deported from Italy* interrogeable à l'aide d'un point d'accès SPARQL (http://lod.xdams.org/sparql).

Cependant, pour des raisons de sécurité, seuls les points d'accès SPARQL ayant été spécifiquement *whitelisted* dans les paramètres de configuration de la Wikibase peuvent être interrogés, cela nécessiterait donc quelques ajustements préalables.

Annexe 12
Service de réconciliation

Pour plus d'infos, consulter cette page (https://linkingthepast.org/about/).

Créer un service de réconciliation basé sur sa propre instance Wikibase

Cette étape vise à explorer comment nous pouvons créer un service de réconciliation OpenRefine* basé sur sa propre instance Wikibase.

*OpenRefine

De nombreux tutoriels documentent l'usage qui peut être fait d'OpenRefine, notamment dans le secteur des GLAM, nous ne reviendrons donc pas là-dessus dans le cadre de ce document. Citons par exemple la documentation (https://msaby.gitlab.io/atelier-openrefine-MASA/index.html) très claire et agréable à consulter, partagée par Mathieu Saby dans le cadre d'un atelier intitulé *Nettoyer et préparer des données avec OpenRefine*

Un service de réconciliation sur mesure

Comme indiqué sur le dépôt GitHub 'openrefine-wikibase' (https://github.com/wetneb/openrefine-wikibase), des adaptations ont été réalisées afin que le service de réconciliation basé sur Wikidata puisse désormais être réutilisé pour n'importe quelle instance Wikibase :

> This service can be configured to run against another Wikibase instance than Wikidata. The Wikibase instance will need to have an associated SPARQL Query Service, and some properties and items will need to be set up. All the relevant values must be configured in the config.py file, and an example of this file for Wikidata is provided in config_wikidata.py.

Après avoir veillé à installer sa propre Wikibase et adapté le fichier de configuration, il est donc possible de lancer son propre service de réconciliation, qui disposera des fonctionnalités suivantes (https://github.com/wetneb/openrefine-wikibase/blob/master/README.md (voir) :

> - Matching columns with Wikibase properties, to improve the fuzzy matching score ;
> - Autocomplete for properties and items ;
> - Support for SPARQL-like property paths such as "P17/P297" ;
> - Language selection ;
> - Reconciliation from sitelinks

Configuration

Le fichier ci-dessous reprend la configuration minimale utilisée pour tester le service de réconciliation basée sur le SPARQL endpoint de notre instance Wikibase.

Pour un usage approfondi, il serait utile d'affiner cette configuration, en partant de l'exemple (https://github.com/wetneb/openrefine-wikibase/blob/master/config_wikidata.py) fourni pour Wikidata.

In [ ]:

```
"""
This file defines a few constants which configure
which Wikibase instance and which property/item ids
should be used
"""

# Endpoint of the MediaWiki API of the Wikibase instance
mediawiki_api_endpoint = 'https://adochs.arch.be/w/api.php' #'https://www.wikidata.org/
w/api.php'

# Regexes and group ids to extracts Qids and Pids from URLs
import re
q_re = re.compile(r'(<?https?://adochs.arch.be/(entity|wiki)/)?(Q[0-9]+)>?')
q_re_group_id = 3
p_re = re.compile(r'(<?https?://adochs.arch.be/(entity/|wiki/Property:))?(P[0-9]+)>?')
p_re_group_id = 3

# Identifier space and schema space exposed to OpenRefine.
# This should match the IRI prefixes used in RDF serialization.
# Note that you should be careful about using http or https there,
# because any variation will break comparisons at various places.
identifier_space = 'https://adochs.arch.be/entity/'
schema_space = 'https://adochs.arch.be/prop/direct/'

# Pattern used to form the URL of a Qid.
# This is only used for viewing so it is fine to use any protocol (therefore, preferabl
y HTTPS if supported)
qid_url_pattern = 'https://adochs.arch.be/wiki/Item:{{id}}'

# By default, filter out any items which are instance
# of a subclass of this class.
# For Wikidata, this is "Wikimedia internal stuff".
# This filters out the disambiguation pages, categories, ...
# Set to None to disable this filter
avoid_items_of_class = None

# Service name exposed at various places,
# mainly in the list of reconciliation services of users
service_name = 'Adochs Reconciliation Service'

# URL (without the trailing slash) where this server runs
this_host = 'http://localhost:8000'

# The default limit on the number of results returned by us
default_num_results = 25

# The maximum number of search results to retrieve from the Wikidata search API
wd_api_max_search_results = 50 # need a bot account to get more

# The matching score above which we should automatically match an item
validation_threshold = 95

# Redis client used for caching at various places
import redis
redis_client = redis.Redis(host='localhost', port=6379, db=0, decode_responses=True)
```

```
# Redis prefix to use in front of all keys
redis_key_prefix = 'openrefine-wikidata:'

# Headers for the HTTP requests made by the tool
headers = {
    'User-Agent':service_name + ' (OpenRefine-Wikibase reconciliation service)',
}

# Previewing settings

# Dimensions of the preview
zoom_ratio = 1.0
preview_height = 500
preview_width = 400

# With which should be requested from Commons for the thumbnail
thumbnail_width = 130

# All properties to use to get an image
image_properties = [
    'P56'
]

# URL pattern to retrieve an image from its filename
image_download_pattern = 'https://upload.wikimedia.org/wikipedia/commons/thumb/%s/%s/%s
/%dpx-%s'

# Fallback URL of the image to use when previewing an item with no image
fallback_image_url = this_host + '/static/cegesoma.png'

# Alt text of the fallback image
fallback_image_alt = 'CegeSoma'

# Autodescribe endpoint to use.
# this is used to generate automatic descriptions from item contents.
# (disable this with: autodescribe_endpoint = None )
autodescribe_endpoint = None #'https://tools.wmflabs.org/autodesc/'

# Property proposal settings

# Default type : entity (Q35120)
default_type_entity = 'Q3617' # = 'Personne' https://adochs.arch.be/wiki/Item:Q3617

# Property to follow to fetch properties for a given type
property_for_this_type_property = 'P13' # ?

# Type expected as target of a given property
subject_item_of_this_property_pid = 'P1629' # ?
```

Lancer le service à l'aide de Docker

Une fois ce document complété, il *suffit* d'installer Docker (https://docs.docker.com/get-docker/) et de lancer les commandes documentées dans le fichier Readme (https://github.com/wetneb/openrefine-wikibase/blob/master/README.md) du dépôt GitHub :

In [ ]:

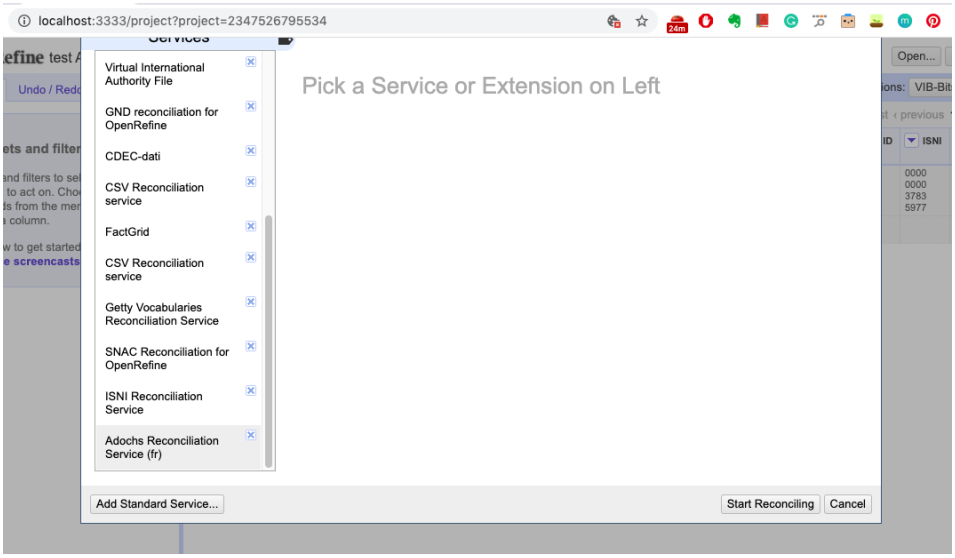
```
docker pull pintoch/openrefine-wikibase
docker run -p 8000:8000 pintoch/openrefine-wikibase
```

Si le service est exécuté en local, le service devrait normalement être accessible à : *http://localhost:8000/en/api (http://localhost:8000/en/api)*

Pour info : le service peut être testé à l'aide du Reconciliation test bench (https://reconciliation-api.github.io/testbench/).

Pour ajouter le service à OpenRefine

- créer ou ouvrir un projet
- cliquer sur l'en-tête d'une colonne
- cliquer (tout en bas) sur Reconcile
- cliquer sur Start reconciling...
- cliquer (tout en bas) sur Add standard service
- sous *Enter the service's URL:* copier l'adresse service de réconciliation associé à la Wikibase (ici http://localhost:8000/en/api (http://localhost:8000/en/api))



Le service peut ensuite être utilisé de la même manière que le service de réconciliation basé sur Wikidata (sous réserve de ce qui a été correctement paramétré bien sûr) :

Cas d'usage

Il est désormais possible de réconcilier des listes de noms de personnes ou de lieux avec les entités stockées dans sa propre instance Wikibase, ce qui élargit les possibles pour une institution confrontée à des doublons, des listes éparses, etc.

En revanche, il n'est pas encore possible de directement injecter des données dans sa Wikibase à partir d'OpenRefine, contrairement à ce que l'outil permet déjà pour Wikidata. Cette demande a toutefois été formulée par des utilisateurs Wikibase et il n'est pas inenvisageable qu'une telle fonctionnalité soit déployée à l'avenir. Pour suivre l'évolution du dossier, voir cette *issue* (https://github.com/OpenRefine/OpenRefine/issues/1640).

NB

À l'heure du test effectué sur l'instance Wikibase 'DataCegeSoma', deux problèmes se sont présentés :

- d'une part, un problème d'accent, qui nécessiterait une investigation plus poussée (voir l'*issue* signalant le problème (https://github.com/wetneb/openrefine-wikibase/issues/82))
- d'autre part, la fonctionnalité de *fuzzy matching* implémentée pour Wikidata, n'a pas marché lors de nos tests, limitant clairement la puissance/l'utilité de l'outil. (À nouveau, une investigation plus poussée devrait être menée pour creuser la question, voir par exemple cette *issue* (https://github.com/wetneb/openrefine-wikibase/issues/80).

Annexe 13
Des métadonnées pour accompagner chaque jeu de données

Cette annexe propose un aperçu d'une initiative mise en place au CegeSoma [158] afin de systématiser l'encodage de métadonnées portant sur les jeux de données eux-mêmes. En l'occurence, ces derniers étant pour l'instant encodés dans des fichiers Excel, il s'agit d'ajouter une feuille Excel visant à préciser le contexte de production du jeu de données, mais également les caractéristiques et règles d'encodage de chaque colonne du fichier, comme le montre la figure 3.

Informations générales	
Titre	Modèle de tableaux de résistants
Auteur(s)	Fabrice Maerten, Anne Chardonnens
Institution	CegeSoma
Email (si extérieur au CegeSoma)	/
Date de création	2020-06-26
Date de clôture	2020-08-20
Provenance	
Issu de	Les données proviennent de dossiers individuels issus du fonds AA 1056,30 (exemple)
Accès	
Nom du fichier	20200727_template_personnes_modele.xls
Emplacement sur le serveur	R:\37_DIROPCESO_CEGESOMA\20_SHARED_DATA_INT\10_ONDERZOEK_RECHERCHE\Tableaux de résistants
Voir également	Inventaire AA1056
Vue d'ensemble	
Nombre colonnes	**30**
Nombre lignes	1000
Personnes	
Caractéristiques communes	Toutes les personnes décrites dans ce fichier sont des résistants décédés ayant agi dans la presse clandestine (exemple)
Descriptif des colonnes	
ID	Identifiant unique pour chaque personne (commence à 1, peut être généré automatiquement)
Numero_inventaire	Numéro de l'inventaire (chiffre arabe ou romain, ou même parfois lettre). En l'absence d'inventaire, laisser vide
Numero_boite	Laisser vide s'il existe un inventaire, sauf s'il s'agit d'une sous-boîte non reprise dans l'inventaire
Nom	Nom de la personne, la particule est placée en premier le cas échéant en respectant la graphie mentionnée (exemples : "De Jongh"
Prenom_officiel	Premier prénom de la personne (exemple : Guilielmus)
Prenom_usuel	Prénom utilisé pour désigner la personne, s'il diffère du prénom officiel (exemple : Guillaume, pour Guilielmus). Si ce n'est pas le
Autres_prenoms	Deuxième, troisième prénoms, etc. s'il y en a ; s'il y a plusieurs valeurs, les séparer par une virgule, avec un espace après la virgule (exemple : "Paul, Fernand")
Genre	Genre de la personne : F pour féminin, M pour masculin (voir feuille *abréviations)*
Langue	Langue(s) parlée(s) par la personne, selon la norme ISO 369-1 : FR pour français, NL pour néerlandais, DE pour allemand, EN pour anglais, etc. (voir feuille *abréviations*) ; s'il y a plusieurs valeurs, les séparer par une virgule (avec un espace après celle-ci)

FIGURE 3 – Exemple de métadonnées destinées à documenter un jeu de données sur des personnes. Source : CegeSoma (document de travail).

158. Cette initiative s'inscrit dans la continuité des recommandations formulées à l'issue de notre étude de cas (section 6.5, p. 255).