Evaluating the success of vocabulary reconciliation for cultural heritage collections

Seth van Hooland[†*], Ruben Verborgh[§], Max De Wilde[†],
Johannes Hercher[◇], Erik Mannens[§], and Rik Van de Walle[§]

[†]Université Libre de Bruxelles
Information and Communication Science Department
Avenue F. D. Roosevelt, 50 CP 123
B-1050 Brussels, Belgium
{svhoolan,madewild}@ulb.ac.be
Phone: +32 2 650 4765

[§]Ghent University – IBBT, ELIS – Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium
{ruben.verborgh,erik.mannens,rik.vandewalle}@ugent.be
Phone: +32 9 33 14959

[◇]Hasso-Plattner-Institute, University of Potsdam
Prof.-Dr.-Helmert-Straße 2–3
D-14482 Potsdam, Germany
johannes.hercher@hpi.uni-potsdam.de
Phone: +49 331 5509 547

May 2012

Abstract

The concept of Linked Data has made its entrance in the cultural heritage sector due to its potential use for the integration of heterogeneous collections and deriving additional value out of existing metadata. However, practitioners and researchers alike need a better understanding of what outcome they can reasonably expect of the reconciliation process between their local metadata and established controlled vocabularies which are already a part of the Linked Data cloud. This paper offers an in-depth analysis of how a locally developed vocabulary can be successfully reconciled with the Library of Congress Subject Headings (LCSH) and the Arts and Architecture Thesaurus (AAT) through the help of a general-purpose tool for interactive data transformation (Google Refine). Issues negatively affecting the reconciliation process are identified and solutions are proposed in order to get a maximum value from existing metadata and controlled vocabularies in an automated manner.

[*]Corresponding author

1 Introduction

1.1 General context and purpose of the paper

The practice of gathering domain and technology experts to debate over a period of years in order to develop and fine-tune metadata schemas and ontologies has lost a lot of its institutional support. The recent eContentplus[1] funding program of the European Commission, for example, explicitly did not fund the development of metadata schemas and the creation of metadata itself (van Hooland et al., 2010). Budget cuts in the US and Europe are forcing Libraries, Archives and Museums (LAM) to adopt a more pragmatic stance regarding metadata creation and management. Funding bodies and grant providers expect short-term results and push cultural heritage institutions to gain more value out of their own existing metadata by linking them to external data sources. It is precisely in this context that the concept of Linked and Open Data (LOD) has gained momentum.

We believe that the semantic enrichment and integration of heterogeneous collections can be facilitated by using subject vocabulary for cross linking between collections, since major classifications and thesauri (*e.g.*, LCSH, AAT, DDC, RAMEAU) have been made available following Linked Data principles. Reusing these established terms for indexing cultural heritage resources represents a big potential for LAM.

Therefore, the paper aims to examine the feasibility of using subject vocabularies as linking hub to the Semantic Web in advance of such effort. Namely, we will consider and answer the following key questions:

- What are the possibilities to reconcile terms from a local controlled vocabulary with well-established vocabularies in an automated manner with the help of a general purpose tool for interactive data transformation?
- What are the characteristics of the reconciled terms and, more specifically, do they offer a sufficient discriminatory value for search and retrieval?

The reconciliation process presents only one step towards the overall ambition of deriving additional value out of existing metadata through Linked Data principles (Heath and Bizer, 2011). After the initial reconciliation of local metadata with one or multiple centralized controlled vocabularies, which is the core focus of this paper, the possibilities and the limitations of search and retrieval across multiple heterogeneous collections need to be assessed. Exciting new possibilities emerge thanks to the links created between distributed collections, but a fair amount of work remains ahead in order to determine how to present and rank results in a meaningful manner for end-users. To an important extent, the success of projects such as Europeana[2] and the Digital Public Library of America[3] will depend on how the value of local metadata can be leveraged through centralized vocabularies and how the links created are presented through meaningful interfaces. This paper presents an analysis of the initial reconciliation step.

1.2 Introducing Linked Data and SKOS

Tim Berners-Lee's TED 2009 mantra "More raw data now!"[4] resonated in the administrative and political domain.[5] The reuse of existing metadata, and the attempt to gain more value out of them by offering publicly available metadata exports, or a direct access over APIs, is little by little unlocking some parts of the silos of metadata built up over decades. The creation and maintenance of highly structured metadata and controlled vocabularies has been the core business of cultural heritage institutions since their professionalization at the end of the 19th century. Accordingly, LAM have gained considerable attention as a valuable source for structured metadata and a growing number of projects, such as Mäkelä et al. (2007), Isaac et al. (2008), Neubert (2009), and Haslhofer and Isaac (2011), are making use of Semantic Web technologies to publish resources from the LAM domain.

The term "Linked Data" is referenced as a set of best practices to publish and connect entities (rather than only documents). The ambition is to create a global data space of networked resources that can be queried with generic tools, rather than putting publishers and agents in charge of understanding custom APIs of so called data-silos.

[1] http://ec.europa.eu/information_society/activities/econtentplus/closedcalls/econtentplus/index_en.htm

[2] http://europeana.eu/

[3] http://dp.la/

[4] Tim Berners-Lee on the next Web, TED Talks, Feb. 2009. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html minute 10:45

[5] See the following projects to illustrate the governmental support regarding the Open and Linked Data movement: http://www.data.gov/, http://www.data.gov.au/, http://opendata-network.org/, http://openbelgium.be/, http://publicdata.eu/.

Basically, this is accomplished by using a set of Web technologies along with generic data formats that extend XML. In this context, the Resource Description Framework (RDF; Brickley and Guha, 2004) and the Ontology Web Language (OWL; Hitzler et al., 2009) make it possible to express meaning of object relations between two resources, or rather to define syntactical constraints that allow an automated derivation of facts. Since reasoning on the web can lead to obfuscating and sometimes funny results (Hogan et al., 2010) the more pragmatic Linked Data approach only aims to provide a uniform model for the access to distributed metadata, by making both the data and its underlying structure (metadata schema) available using persistent and HTTP dereferenceable URIs (Sauermann and Cyganiak, 2008). Therefore, datasets that are encoded in RDF/OWL can be easily queried using the RDF Query Language (SPARQL; Prud'hommeaux and Seaborne, 2008), independently of the metadata schema, or customized encoding structures in local repositories. Consequently the data can be exposed using so-called SPARQL endpoints or triple stores to ease the interlinkage of distributed resources. We refer to Heath and Bizer (2011) for a state-of-the-art overview of Semantic Web technology and best practices.

In the context of making authority files and subject vocabularies part of the *Linked Data cloud*[6], the Simple Knowledge Organizing System (SKOS; Miles and Bechhofer, 2009) has gained considerable popularity. SKOS was developed as an RDF Schema to represent thesauri in the Semantic Web (Alistair et al., 2005, p.17). It is compatible with relevant standards, such as ISO 2788:1985/1986 and ANSI/NISO Z39.19, and has been preferred over other formats, such as Topic Maps[7] or zThes[8], because of its flexible structure and standardization by the W3C (Pastor-Sanchez et al., 2009).

SKOS provides straightforward relation types to express terminological structure of subject vocabularies. The properties `skos:prefLabel` and `skos:altLabel` are used to describe synonymous denotations of one concept, whereas `skos:narrower`, `skos:broader`, and `skos:related` describe *hierarchical* and *associative* relationships between two concepts within one thesaurus. Additionally, SKOS is designed to match descriptors of different vocabularies by assigning `skos:closeMatch`, `skos:exactMatch` or `skos:narrowMatch` as properties between two concepts (Isaac et al., 2008). As such, SKOS provides a very generic approach to the representation of authority data at large and leaves aside the specificities of pre-coordinated systems, which concatenate independent terms in a specific order. To support the individual components of pre-coordinated subject labels, the Metadata Authority Description Schema in RDF (MADS/RDF) has been proposed by the Library of Congress (of Congress, 2011). Since its publication in 2010, the MADS/RDF initiative has attracted a fair amount of criticism. Commentators suggest to give priority to well-known issues with the syndetic structure of LCSH due to the errors in the automated process, which converted undifferentiated `SeeAlso` relationships to `Broader`, `Narrower`, or `Related` relationships in 1987, before making the error-prone semantics available as Linked Data (Spero, 2008). More importantly, Linked Data practitioners criticized the fact that the data model issued its own classes and properties without reaching out to existing ontologies.[9]

1.3 Related work regarding the reconciliation of metadata and vocabularies

The shift from printed books to digital tools for the management and use of controlled vocabularies already lead in the 1990s to a considerable body of research regarding automated and semi-automated methods for achieving interoperability between vocabularies (Doerr, 2001). Isaac et al. (2008) identified four general approaches towards vocabulary reconciliation or alignment: 1) lexical alignment techniques, 2) structural alignment, 3) extensional alignment, and 4) alignment using background knowledge. We decided to focus on lexical alignment technologies, as most of the terms can be reconciled by taking care of lemmatization, harnessing preferred labels or computing string similarity (*cf.* 4.2.2).

Lexical alignment can be performed easily by vocabulary managers during the reconciliation process using an Interactive Data Transformation (IDT) tool such as *Google Refine*, as will be demonstrated throughout the paper. Tools of this type have been deployed in the context of large-scale Linked and Open Data projects such as data.gov.uk[10] and the "Dollars for Docs" investigative studies by ProPublica.[11] Within the specific context of cultural heritage metadata,

[6] Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch at `http://lod-cloud.net/`

[7] `http://www.isotopicmaps.org/`

[8] `http://zthes.z3950.org/`

[9] Jeffrey Beal has presented on overview of different comments formulated regarding the MADS/RDF initiative, available on `http://metadata.posterous.com/lcs-madsrdf-ontology-and-the-future-of-the-se/`.

[10] `http://data.gov.uk/`

[11] `http://www.propublica.org/series/dollars-for-docs`

Allinson (2012, p. 46) mentions the use of Google Refine and the RDF extension within the OpenART project[12] to "map data in spreadsheets against our ontologies". Despite the fundamental importance of the reconciliation procedure for a project with the ambition to interlink local metadata with external vocabularies and other metadata collections, no information is given regarding the success rate of the reconciliation and what specific steps were involved. The article only briefly mentions the need for some supplementary scripting and data clean-up. These are exactly the issues this paper addresses in detail, as they are fundamental for the understanding of the processes and methodologies needed to successfully bring metadata into the Linked Data cloud. Fadi Maali and Peristeras (2011) analyze several technical approaches to the implementation of reconciliation services that would facilitate the linking process between third parties and well-established LOD hubs. Performance measurement of different services providing domain-independent support for RDF interlinking lies at the heart of the paper, but it offers a thorough description of the reconciliation workflow of the Google Refine RDF extension[13].

The reconciliation of a subject vocabulary is a crucial task for a wide range of applications in LAM institutions and has been tackled before in various projects (Tudhope et al., 2011; van der Meij et al., 2010; van Erp et al., 2011). Van Erp et al. (2011) demonstrate how to achieve a mapping between name-authority-thesauri using sophisticated NLP technology. Wikipedia articles of historic events are examined for significant phrases that indicate relationships between an entity (person, location, time) and an event. Entities, which occur in the same context and in different authorities at the same time, are used to determine a mapping between two thesauri classes. This method is only usable for named entities (people, places, organizations etc.) and did not work out for the subject vocabulary we intended to use. Furthermore, our approach differs because we chose to explore *non-expert* software, which can be mastered and handled by vocabulary managers who do not necessarily have advanced skills in computer science or language technology.

In the scope of the CATCH/STITCH project, van der Meij et al. (2010) developed a service[14] to perform semi-automated vocabulary alignment between large vocabularies, such as RAMEAU, ICONCLASS, and LCSH. Several vocabularies from the cultural heritage domain are imported, matched, and published using SKOS matching relations. Attention is paid to the evaluation of machine created mappings between interlingual vocabularies derived from the MACS Project (Landry, 2009), leading to a hosting and alignment service for libraries that do not have the resources or the competencies to provide such infrastructure. However, performance and usability issues are largely neglected. Another drawback of van der Meij et al. (2010) is that vocabularies need to be transformed into RDF/SKOS first, whereas we can export our reconciliation results directly into RDF/SKOS format, which consequently can be exported into arbitrary triple stores for reuse.

Within the STELLAR[15] project, the CIDOC Conceptual Reference Model[16] was populated with mappings to glossaries and thesauri to support indexing and search within archeological resources. Tudhope *et al.* also provide a conversion service to transform CSV files into RDF/SKOS, but the scope of the STELLAR project is mostly limited to the archeological domain and its tools are not generalizable and as straightforward to use as Google Refine.

Crowdsourcing approaches have been used to encourage patrons in producing user-generated metadata (van Hooland et al., 2011). The linking between a folksonomy, captured from Delicious, and the LCSH[17] has been studied by Yi and Chan (2009). The authors selected three sets of 100 tags, based on their frequency (high, mid, and low frequency) and analyzed the possibilities of automatically linking the tags with LCSH. Each user tag was compared on the basis of "complete-word matching", meaning that a tag is linked with a subject heading when the tag "appears in the heading as a complete word" (Yi and Chan, 2009, p. 887). As the authors mention, this means that the tag "computer" results in a match with the subject heading "computer networks" but the tag "network" is not considered to be a complete-word match with the same heading as the tag is only a portion ("network" for "networks") of the word in the subject heading, and "not as a complete word" (Yi and Chan, 2009, p. 887). This approach resulted in 60.9 percent of matches. Our reconciliation process differs significantly from Yi and Chan's approach as we perform the reconciliation 1) on the entire corpus and not on samples of statistically insignificant size, 2) on the basis of a complete match between a locally developed vocabulary and the LCSH.

To sum up, we depart from previous work by empirically demonstrating the feasibility, for cultural heritage practitioners, to reconcile local vocabularies with well-established controlled vocabularies through the help of a general purpose tool for interactive data transformation. This claim is supported through the use of a representative case study of which the data, the tools and the documentation of performed actions are freely available online, allowing others

[12] https://yorkdl.wordpress.com/tag/openart-jisc-rdtf/
[13] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/
[14] http://www.cs.vu.nl/STITCH/repository/
[15] http://hypermedia.research.glam.ac.uk/kos/stellar/
[16] http://cidoc-crm.org/
[17] http://id.loc.gov

to repeat and verify our research results. The parameters impacting the reconciliation process are described in detail and, where possible ,semi-automated solutions are developed to solve encountered problems. Existing work fails to discuss these issues in detail and does not offer concrete examples of what parameters impact reconciliation positively and negatively. We therefore believe our paper is a helpful contribution to catalyze experimentation with integrated and topic-oriented navigation among heterogeneous collections within the cultural heritage sector.

1.4 Structure of the paper

The remainder of the paper is structured as follows. Section 2 presents our methodology, which is centered on a case study. The used metadata, the controlled vocabularies that are already connected to the Linked Data cloud, and the interactive data transformation tool used for the reconciliation process are described in detail. Before the reconciliation process can take place, we need to identify potential metadata quality issues that could lower the success rate of the reconciliation. The profiling and cleaning of the metadata is therefore described in section 3. The reconciliation process itself is then explained in section 4, which presents the results of the out-of-the-box reconciliation with the LCSH and AAT vocabularies through the help of the RDF extension for Google Refine. This section also shows how the results can be significantly augmented by a set of simple and automated manipulations. Section 5 proposes an in-depth analysis of the results by studying the formal and semantic characteristics of the reconciled terms, based upon manual analyses of samples, and an automated approach to evaluate the granularity of the reconciled terms. We then return to the two initial research questions in the conclusions and sum up our findings.

2 Methodology

Recent efforts, such as the W3C Library Linked Data Incubator Group[18], the International Linked Open Data in Libraries, Archives and Museums Summit (LOD-LAM)[19], and ResearchSpace[20] are very encouraging. The Linked Data Pilot of Europeana[21] holds the potential of large groundings for interlinkage of LAM, since the metadata of more than 17,000,000 digitized cultural heritage objects have been published following the Linked Data principles (Haslhofer and Isaac, 2011). These initiatives are essential, but there is a need for more experimentation with and uptake of Linked Data among collection holders who do not have the resources of large-scale international projects.

This paper therefore wants to demonstrate the feasibility for collection holders and researchers to reconcile vocabularies with the help of general purpose interactive data transformation tools. In order to support this statement, the paper presents an extensive case study focusing on a detailed description of the reconciliation process and an explanation of the stumbling blocks one might encounter both from the point of view of the local metadata and the centralized vocabularies. Throughout the different steps of the case study, the paper presents the type of results which can be reasonably expected of the initial "out-of-the-box" reconciliation and the enhanced process.

It is outside the scope of the article to provide a fully detailed step-by-step description of the profiling, cleansing and reconciliation process, but a "Free your metadata"[22] website has been created as a side-project, offering in-depth documentation and screencasts on how to repeat the cleansing operations and the reconciliation process with the LCSH, as described in the paper. The next sections will present the different elements of the case study.

2.1 Metadata: export from the Powerhouse museum

The Powerhouse Museum in Sydney provides a freely available metadata export of its collection on its website.[23] The museum is one of the largest science and technology museums worldwide, providing access to almost 90,000 objects, ranging from steam engines to fine glassware and from haute couture to computer chips. The Powerhouse has been very actively disclosing its collection online and making most of its data freely available. From the museum website, a tab-separated text file can be downloaded. The unzipped file (58MB) contains basic metadata (17 fields) for 75,823 objects, released under a Creative Commons Attribution Share Alike (CCASA) license.

[18] http://www.w3.org/2005/Incubator/lld/
[19] http://lod-lam.net/
[20] http://www.researchspace.org/
[21] http://data.europeana.eu/
[22] http://freeyourmetadata.org/.
[23] http://www.powerhousemuseum.com/collection/database/download.php

The reconciliation process specifically focuses on the *Categories* field, which is populated with terms from the Powerhouse museum Object Names Thesaurus (PONT).[24] This thesaurus was created by the museum and first published in 1995. As of April 2012, the thesaurus consists of 6,504 preferred terms and 2,091 non-preferred terms. The controlled vocabulary is currently available as a downloadable PDF file[25], but an online thesaurus browser should be published shortly on the museum's website. PONT recognizes Australian usage and spelling, and reflects in a very direct manner the specificity of the collection. This results, for instance, in a better representation of social history and decorative arts, whereas only a minimal number of object names exist in the domains of fine arts and natural history. According to Sebastian Chan, Head of Digital, Social and Emerging Technologies at the Powerhouse museum, the staff of the museum responsible for the *Categories* field are trained registrars, whereas curators provide the metadata regarding significance, history, and provenance of the collection.

2.2 Established controlled vocabularies available as Linked Data

To ease the linking of resources between collections, common vocabularies are needed in order to identify shared concepts. A limited number of controlled vocabularies have a broad and international uptake within the cultural heritage sector. We decided to make use of the two best-known and most widely-used controlled vocabularies within this sector, which are already available (optimally or suboptimally) as Linked Data: the LCSH and the AAT. These two vocabularies differ greatly in their principles and usage, as the LCSH is primarily pre-coordinated, whereas the AAT is post-coordinated. Pre-coordination is a "combination of concepts, classes or terms of a controlled vocabulary at the time of its construction or at the time of using it for indexing or classification" and post-coordination refers to a "combination of preferred terms of a controlled vocabulary at the time of searching" (ISO25964-1, p. 9). Both approaches have their advantages and disadvantages, as described by Bodoff and Kambil (1997) and Library of Congress Cataloging Policy and Support Office (2007). In the specific context of this paper, abstraction of the pre-coordinate aspect of the LCSH is made to a significant extent, since the reconciliation process focuses on either main headings or subdivisions. In other words, no use is made of the complex headings that combine both (*e.g.*, *"Washington, George, 1732-1799–Numismatics"*). However, this can even take place on the level of unique headings pre-coordination, as is the case for example in the heading *"Mass media and children"*, which is composed of the two individual terms *"mass media"* and *"children"*. Throughout the paper, we will use the denominator *term* to designate the representation of concepts within the AAT and the PONT, and the denominator *heading* to refer to the representation of concepts within the LCSH.

2.2.1 LCSH

The LCSH is not only the most adopted subject indexing language worldwide (Anderson and Hofmann, 2006), but as (O'Neill and Chan, 2003) mention, they are also the largest available controlled vocabulary in English. The LCSH has initially been developed for bibliographic descriptions covering a large variety of subject areas. As mentioned above, it is primarily a pre-coordinated system but, when confronted with a complex subject for which no single heading exists, a post-coordinated approach may be taken by cataloguers in order to reach a sufficient level of specificity (Chan, 2005; Library of Congress Cataloging Policy and Support Office, 2007).

The SKOS-ification of the LCSH has been well-documented (Summers et al., 2008). The use of LCSH also allows a future mapping to the Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié (RAMEAU) and SchlagWortnormDatei (SWD), the French and German indexing subject headings list, giving potential access to resources in those languages as well. In 1998, four European national libraries established the Multilingual Access to Subjects (MACS) project in the context of the Conference of European National Librarians (CENL). The manual alignment process between the RAMEAU, the LCSH, and the SWD took ten years and is now implemented at the Swiss National Library and the Deutsche Nationalbibliothek (Landry, 2009). The STITCH project has reused the outcomes of the MACS project and worked on vocabulary alignment by automatically detecting inter-vocabulary semantic mappings between vocabularies such as Iconclass, the Brinkman thesaurus, and WordNet (van der Meij et al., 2010). The interlinking of these vocabularies in multiple languages has the potential to facilitate multilingual and cross-collection search and retrieval.[26]

The LCSH has been formalized in RDF using the SKOS vocabulary, which enables us to reconcile the Powerhouse Museum metadata against a Linked Data source. As of April 2012, they consist of 407,099 concepts, all of which have a

[24] `www.powerhousemuseum.com/collection/database/thesaurus.php`

[25] `http://www.powerhousemuseum.com/pdf/publications/phm-thesaurus-sept09.pdf`

[26] For a concrete view on the scope of the project, consult the CATCH Vocabulary and alignment repository demonstrator available at `http://www.cs.vu.nl/STITCH/repository/`.

preferred label (`skos:prefLabel`) and some having one or multiple *alternate* labels (337,841 `skos:altLabels` in total). During the study, we noticed that the RDF constructs chosen by the Library of Congress to express the LCSH are not definitive yet. For instance, the way in which subdivisions are formalized changed between 2011 and 2012, necessitating configurable processing tools. The LCSH dataset can be queried through a SPARQL endpoint,[27] which unfortunately suffers from availability and performance problems. Alternatively, a serialized RDF version of the dataset is available on the Library of Congress website,[28] which we used to set up a private endpoint dedicated to the reconciliation tasks, eliminating the performance issues mentioned above.

2.2.2 AAT

The AAT is the "most widely known specialist thesaurus" (Broughton, 2006, p. 41), developed for the cultural heritage domain with a specific focus on art, architecture and material culture. Constructed and maintained by the Getty Foundation, the AAT is used for the description of works about art and works of art. AAT terms can therefore act both as cataloging terms for books but also, more importantly, as descriptors of physical characteristics of objects, resulting in the presence of terms such as "cracks" and "color shift". The history of the development of the AAT is very much intertwined with that of the LCSH. Dissatisfaction with the use of the LCSH by art librarians has been a main impetus for the development of the AAT in the 1980s, as the coverage of the arts and architecture by the LSCH was deemed insufficient. Archives and museums at the time either developed in-house specific vocabularies or did not have any controlled subject access. The need for a thesaurus for the cultural heritage domain was expressed in 1979, and work on the AAT took place throughout the 1980s (Petersen, 1990). Within the context of rising computerized cataloging and indexing, specific importance was given to a rigorous and consistent hierarchical approach, allowing users to browse through nested terms and giving indexers the freedom to combine single concepts. Despite the critiques regarding the pre-coordinated approach of the LCSH and the inconsistencies within its syndetic structure, the AAT drew on the LCSH during the phase of terminology gathering "because of LCSH's long-term preeminence as an indexing vocabulary" (Petersen, 1990, p. 648). The website of the Getty provides a clear overview of all other sources used for terminology gathering.[29] As a thesaurus, the AAT is generally used for post-coordinate indexing. However, pre-coordinated compound concepts can be included within a thesaurus when a compound concept semantically refers to the union of two or more sets of documents rather than their intersection (Will, 2012). Furthermore, the AAT has its own specific rules for the encoding of compound concepts (Soergel, 1995).

Similarly to the LCSH, the AAT is available online through a Web interface.[29] The AAT interface, however, targets human users and does therefore not include a SPARQL endpoint. While the full contents of LCSH are freely available for download, the Getty Research Institute requires a license to download the AAT for offline use. The downloadable version is contained in a proprietary XML format, which required conversion to RDF, using an ontology such as SKOS. As of April 2012, the converted version of the AAT consist of 34,800 concepts, all of which have a *preferred* label (`skos:prefLabel`) and in total 81.028 *alternate* labels (`skos:altLabel`). The AAT license does not permit exposure of this RDF version, but we provide the conversion script to allow verification of our experiments.[30]

2.3 General purpose tool for interactive data transformation: Google Refine

Our previous research regarding data profiling of museum metadata relied on an open-source general-purpose data profiling tool.[31] Despite its merits, this profiler has some limitations: only five analyses are available, an XML profile specification file has to be manually set up, and the tool is used through the command line.

Several general-purpose tools for interactive data transformation have been developed over the last years, such as Potter's Wheel ABC[32] and Wrangler[33]. Guo et al. (2011) provide a good overview of recent developments in this field and also propose automated strategies to augment the usability of data for downstream tools. Google Refine[34] (formerly Freebase Gridworks) has recently gained a lot of popularity and is rapidly becoming the tool of choice to quickly and efficiently process, clean, and enrich large amounts of data in a browser based interface. Made available in October 2010,

[27] `http://api.talis.com/stores/lcsh-info/services/sparql`
[28] `http://id.loc.gov/download/` choose "LCSH RDF/XML" or "LCSH N-Triples"
[29] `http://www.getty.edu/research/tools/vocabularies/aat/`.
[30] `https://github.com/RubenVerborgh/Vocabulary-Processing`
[31] The data profiler developed by Yves Bontemps is available at `http://sourceforge.net/projects/dataprofiler/`.
[32] `http://control.cs.berkeley.edu/abc/`.
[33] `http://vis.stanford.edu/papers/wrangler/`.
[34] `http://code.google.com/p/google-refine/`

it provides a number of analyses such as splitting or joining multi-valued cells, converting data into new forms, facetting textual or numerical values, detecting blank cells, and trimming whitespace. The tool also offers a powerful clustering functionality, based on the key collision and nearest neighbor algorithms, allowing to detect near-duplicates. Google Refine further allows to reconcile data with existing knowledge bases, creating the connection with the Linked Data vision.

The DERI research group has developed an RDF extension for Google Refine, which can be downloaded for free.[35] The RDF extension allows users to add SPARQL endpoints to the reconciliation process. DBpedia is for example added, so that the content of the categories field can be matched to terms described as SKOS concepts in DBpedia. More specialized sources such as the LCSH and the AAT can also be used, as we will show later in this paper.

3 First step: profiling and cleansing of metadata

3.1 Context of data profiling and cleansing

Data quality is obviously not a new issue for the Library and Information Science field. In the library context, managers have been put under pressure since the 1980s to cut back on human resources attributed to cataloging, leading – especially in the US – to a vivid debate regarding what exactly quality cataloging means. The work of David Bade (2009) for instance, provides valuable insights into this discussion.

Metadata practitioners who worked on aggregation projects, harvesting metadata from different partners, must acknowledge that the quality of existing metadata is hardly questioned and only becomes visible once they are put to work and queried by a large number of users. After all, what collection holder wants to stand up in the middle of his or her peers and warn them about the low quality of his or her metadata? This misplaced trust causes delays and failures when metadata do not live up to expectations (van Hooland et al., 2008). More importantly, we have to acknowledge there are no established methodologies or tools for metadata quality evaluation. To put it more bluntly in the words of Diane Hillmann: "There are no metadata police" (Hillmann and Phipps, 2007), even with initiatives such as the Pedantic Web group[36], offering a typology of errors to be avoided when publishing metadata in the context of the Semantic Web. Conceptual frameworks on metadata quality criteria have been developed, but both practitioners and researchers hardly apply these theoretical frameworks to analyze their metadata.

In the absence of concrete methodologies and tools, metadata practitioners usually believe that producing information describing the quality of their metadata is too big a step to be taken. This paper will therefore present the notion of data profiling, defined by Olson (2003) as *"the use of analytical techniques to discover the true structure, content, and quality of a collection of data"*. Before asking the question of how to link metadata from different sources, we need to develop strategies to check their initial quality and possibly solve issues which might disturb the reconciliation process amongst different resources. Only then will we be able to evaluate the real added value of the Linked Data approach for the cultural heritage sector. We will illustrate now, with the help of Google Refine, how a quick overview of the metadata quality of a collection can be obtained and what types of cleansing actions should be taken.

3.2 Profiling and cleansing the Powerhouse Museum metadata

Prior to presenting the results obtained from the profiling of the Powerhouse Museum metadata, we clearly want to state that we are not pointing fingers at issues from this particular collection. Instead, we want to give a honest appraisal of common issues, which can help other collection holders to act upon their metadata.

3.2.1 Deduplication

After loading the metadata into the application, the first operation we need to perform is to detect and remove duplicates. This can easily be performed by sorting on *Record ID* and performing the *Blank down* command, detecting consecutive duplicated cells. In this manner, 86 records were identified and deleted from the metadata set. The sorting operation also allows to detect three records which did not contain a record ID or any other information, except for an invalid persistent link, which was automatically generated. These three records were also deleted.

[35] `http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/`
[36] `http://pedantic-web.org/`

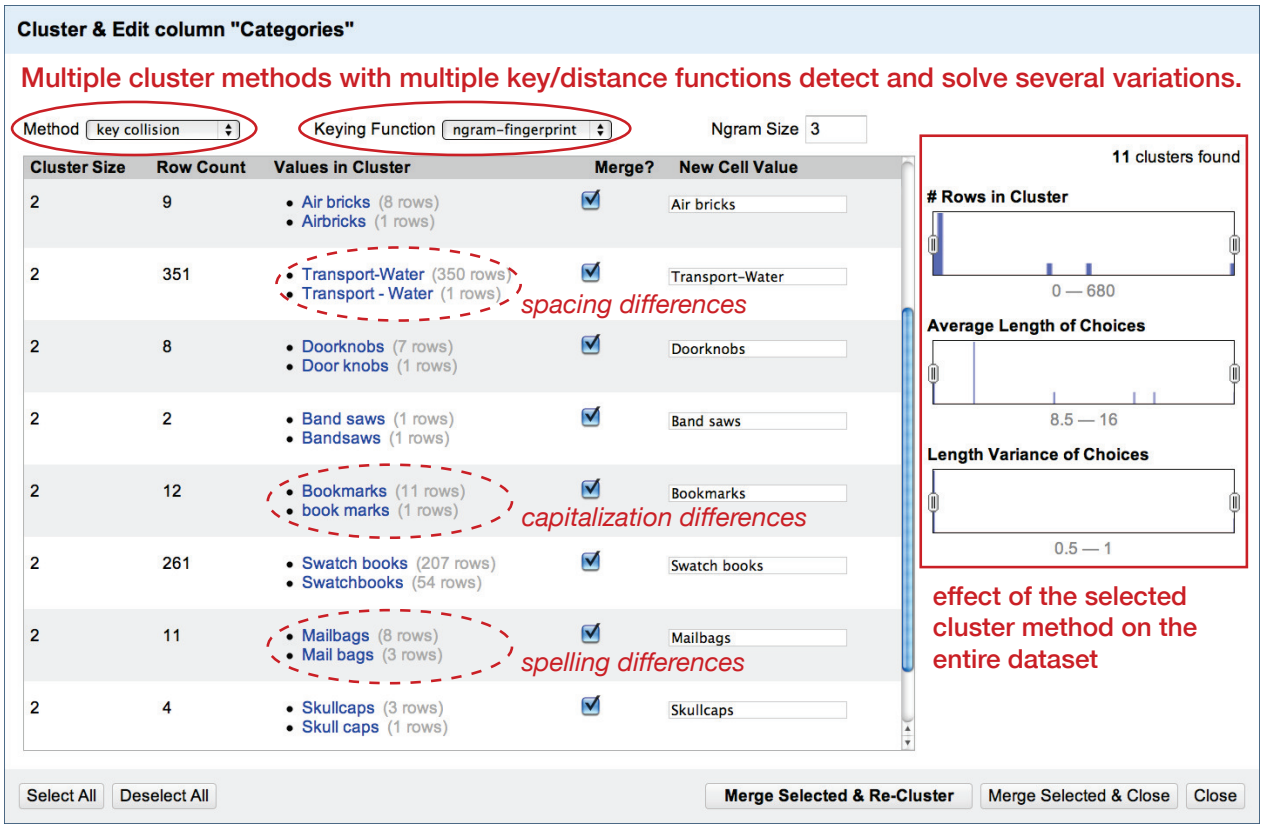

Figure 1: Clustering allows the detection of terms with inconsistent spelling.

3.2.2 Atomization

Once the duplicate records have been eliminated, we can have a closer look at the *Categories* field. A quick glance at an example such as *"Abacus|Writing equipment|Calculating Instruments|Writing and Printing Equipment"* illustrates one of the biggest hurdles for automated metadata analysis: field overloading. On average, 2.25 categories are attributed per object and are contained within the same field. These values mostly go from the most specific to the most general heading. However, a lot of exceptions make it impossible to presume that we could automatically extract the most specific term by taking the first value appearing within the field. Outright doubles also occur, as in *"Photographs|Booklets|Documents|Photographs"*: 1,668 records (about 2%) are concerned with double keyword entries. In order to analyze in detail the use of the keywords, the values of the *Categories* field need to be split out in individual cells on the basis of the pipe character (" | "), expanding the 75,823 records into 170,311 rows.

3.2.3 Applying facets and clustering

Once the content of a field has been properly atomized, filters, facets, and clusters can be applied to give a quick and straightforward overview of classic formal metadata issues. By applying the custom facet "facet by blank", one can immediately identify the 461 records that do not have a category, representing 0.01% of the collection. Applying a text facet to the *Categories* field allows an overview of the 4,893 different categories used in the collection. The headings can be sorted alphabetically or by frequency, giving a list of the most-used terms to index the collection. The top three headings are *"Numismatics"* (8,012), *"Ceramics"* (7,389) and *"Clothing and dress"* (7,280). After the application of a facet, Google Refine proposes to cluster facet choices together based on various similarity methods, such as nearest neighbor or key-collision. As Figure 1 illustrates, the clustering allows to solve issues regarding case inconsistencies, incoherent use of either the singular or plural form, and simple spelling mistakes. Google Refine presents the two related values and proposes a merge into the more recurrent value of the two.

Once the clustering process cleaned up the list of terms, the list can easily be exported as tab-separated values to

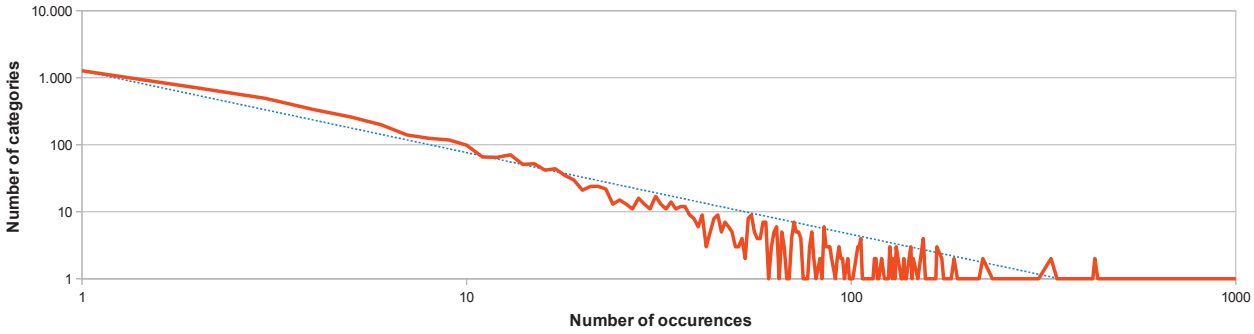

Figure 2: Most terms are only used a few times, resulting in a positively skewed distribution *(shown in a log-log plot)*.

make, for example, visualizations such as a distribution graph. The result is a positively skewed distribution (Figure 2): the top 20 categories are used to describe 88% of the objects, 1,272 categories (26%) are used only once in the collection, and 3,654 categories (75%) are used less than 10 times.

4 Second step: reconciliation of metadata

4.1 Context of reconciliation

For long, linguists and computer scientists have been trying to construct a comprehensive ontology of the world, enabling automated reasoning tasks on human-structured data. Vital to the success of such an ontology are the number and nature of the relationships between different concepts. In recognition of the importance of those relationships, the Semantic Web community formed a Linked Data movement (Bizer et al., 2009), which strives to publish interlinked data in a structured format.

Berners-Lee (2006) has put forward a five-star scheme against which the quality of the data made available on the Web can be evaluated. The highest level is reached when data is linked against other sources, implying the use of well-defined relationships such as equivalence, inclusion, inheritance, *etc*. As a result, machines are able to both broaden and deepen their understanding of data, since links provide the possibility to look up new data and to relate uninterpreted data to well-understood concepts.

Similarly, reconciliation is the process during which we map metadata concepts in a certain (often situation-specific) vocabulary to another (often more commonly used) vocabulary. In case the latter vocabulary is part of the Semantic Web, this reconciliation act paves the path for the fifth star in the Linked Data scheme, as it provides the links that will annex the metadata to the Linked Data cloud. Subsequently, machines will be able to access and interpret these metadata, based on previously acquired knowledge. Reconciliation therefore plays a crucial role in the public availability and dissemination of metadata.

4.2 Reconciling the Powerhouse Museum metadata with the LCSH

4.2.1 Initial reconciliation

When discussing reconciliation results, we will distinguish between the number of matched *rows* and the number of matched *records*. In total, 167,004 categories have been assigned to 75,275 objects, giving 2.2 categories per object on average. Google Refine presents each category assignment as a row and each object as a record. A successful automatic reconciliation of a category assignment is called a *matching row*, whereas we consider a record to be a *matching record* when at least one of its assigned categories has been reconciled.

We first set up our LCSH endpoint in Google Refine by registering it as a SPARQL reconciliation service. Reconciliation is started on the *Categories* field, which has been previously cleansed, clustered, and split into individual values. Initially, 20,300 rows (12.2%) or 19,157 records (25.5%) were matched, if only preferred labels are considered. When alternate labels were also taken into account, match scores went up to 56,501 rows (33.4%) or 42,268 records (56.2%).

Within the Google Refine interface, the unmatched rows fall into two categories: those for which one or several suggestions where identified, and those that do not relate to any LCSH heading. A suggestion consists of a heading label and a score indicating the estimated correspondence between the heading and the content of the row category. Again, two cases occur: either only partial matches exist, or multiple exact correspondences were found (since a single exact correspondence would have caused a match).

The former case includes slight formal variations such as *"Ice-skates"* (PHM) versus *"Ice skates"* (LCSH) or categories that only match with one part of a multiple-word heading such as *"Leaflets"* (PHM) versus *"Leaflets dropped from aircraft"* (LCSH). The latter case of multiple exact correspondences originates in the fact that different types of headings can share labels. In complement with the main headings, four types of subdivisions exist in the LCSH: topical, geographic, chronological and form (Chan, 2005). One prominent example is *"Numismatics"* (affecting 8,012 rows), which exists in the LCSH both as a main heading (`sh85093255`) and as a subdivision (`sh99005172`). The latter is used as a complement to another heading (*e.g.*, *"Washington, George, 1732-1799–Numismatics"*). Obviously, the categories of the Powerhouse museum should match with the main headings, but there is no way for Google Refine to reckon with this preference.

The impact of *term qualifiers* on the reconciliation success rate also has to be investigated. The use of qualifiers is an established practice within thesauri in order to address the issue of homonyms, by disambiguating a term through the inclusion of one or more words enclosed within parentheses following the term. Qualifiers may also be used to facilitate the understanding of an obscure term. Their use can make the application of a thesaurus more cumbersome and may cause problems within automated systems. Therefore the ISO states that "their use (especially in preferred terms) should be avoided. Multi-word term should be preferred to a single-word term with a qualifier, as long as the compound form occurs in natural language" (ISO25964-1, p. 22). However, the ISO also acknowledges that it is "often difficult and subjective" (ISO25964-1, p. 40) to decide whether to include a compound term or not, so not opting for a qualifier often results in equally complex discussions.

Throughout the collection of the Powerhouse Museum, only 101 records - representing 0.01% of all records - are described with terms containing a qualifier. Out of these records, only one term with a qualifier, *"Blowpipes (Weapons)"* was reconciled with the LCSH (`sh85015049`). As the presence of qualifiers within the categories is almost negligible, the cost/benefit relation does not justify the development of automated actions to potentially enhance the reconciliation success. Throughout the LCSH itself, 24.6% of the terms (consisting of both preferred labels and alternate labels) make use of qualifiers. Over time, qualifiers have not only been used within the LCSH to disambiguate homonyms or to clarify obscure or foreign terms, but also to render a heading more specific (*e.g.*, *"Olympic games (Ancient)"*), to specify the genre of a proper name (*e.g.*, *"Banabans (Kiribati people)"*), and to indicate the medium used to perform music (*e.g.*, *"Concertos (Violin)"*) (examples taken from Chan, 2005, p. 53-54). Svenonius points out that "the use of a single device for more than one function can be problematic, particularly in times of technological change" (Svenonius, 2000, p. 21), as the inconsistent use of qualifiers makes potential post-hoc automated processing complex. There is not only a large variation in the functions the qualifiers perform, there is also a considerable variation in the amount of words used for a qualifier. ISO states that "qualifiers should be as brief as possible, ideally consisting of one word" (ISO25964-1, p. 22) but only 51.5% of the LCSH qualifiers consist of one word. 33.5% are made up of two and 10.7% of three words, the rest being spread out from four to eleven words.

4.2.2 Enhanced reconciliation

Confronted with the issues described above, several approaches were tested to augment the reconciliation process by automated processing, both on the level of the Powerhouse Museum categories and the LCSH. The approaches are presented in the order of impact they have on the reconciliation success.

First of all, a solution was sought to tackle the problem of multiple exact correspondences due to the fact that several main headings and subdivisions share the same label (1,269 cases in total): preference needed to be given to main headings over subdivisions. To achieve this, the the LCSH RDF dataset was preprocessed to retain only one heading per unique preferred label. Alternate labels were added only to the extent that did not cause clashes with other labels. The resulting data set was fed into an internal SPARQL endpoint. As a consequence, the success rate of the subsequent reconciliation substantially increased: 78,868 rows (47.2%) or 58,840 records (78.1%) were matched.

Looking at the remaining mismatches, the issue arose whether terms are represented in singular or plural form. Preferences regarding the use of singulars or plurals in controlled vocabularies vary in between languages and cultures.[37]

[37] In French and German, the singular form is preferred in thesauri in order to conform to the rules of use of a dictionary. In English and Spanish, the choice is based on whether a term is a count or non-count noun. Within the context of a multi-lingual thesaurus, such as the AAT for example, the

As a general rule, LCSH expresses concepts in the singular (*e.g.*, *"Viscosity"*) and objects in plural (*e.g.*, *"Dogs")*, as indicated by Broughton (2004). However, a multitude of exceptions exist, partly described by Chan (2005). The categories assigned to the Powerhouse Museum objects are always in plural.

Therefore, an additional reconciliation on the singularized versions of the categories was performed. The analysis of the additional matches reflects the inconsistency of how the LCSH handles singular and plural forms but, more importantly, also the inconsistent use of alternate labels to include variant forms into the vocabulary. The heading *"Dogs"* (`sh85038796`)—in plural because it is a domestic animal, whereas biological species are generally in singular form (Chan, 2005)—has the altLabel *"Dog"*, but *"Stone"* (`sh85128287`) does not have the altLabel *"Stones"*, resulting in a mismatch of 55 records of the Powerhouse.

In total, stemming allowed to match 3,335 supplementary terms. However, it is important to note that this approach can be problematic when semantic differences exist between the singular and the plural form of a term. For example, 166 rows with the term *"Sculptures""* from the Powerhouse museum have been matched with *"Sculpture"*(`sh85119004`). Within the thesaurus of the Powerhouse museum, the term expresses the object, as it has as broader term *"Artworks"*, as narrower terms *"Busts"*, *"Ceramic forms"*, *"Maquettes"*. Within LCSH, *"Sculpture"* mainly expresses the technique, with narrower terms such *"Bamboo carving"* and *"Cement sculpture"* for example, but other narrower terms such as *"Monuments"* and *"Altered sculptures"* refer more to the object, which is also implied by the plural form of the terms.

The previous section mentioned the presence of qualifiers within the LCSH and discussed the varying manner in which qualifiers are used throughout the LCSH. We experimented with the omission of qualifier information from the reconciliation terms, in the sense that the string representing the heading used for the reconciliation algorithm did not contain the qualifier. This approach was developed in order to automatically reconcile headings containing a qualifier, despite the fact that no identical heading with another qualifier exists. For example, the heading *"Bayot (African people)"* has the qualifier *"(African people)"* to specify the genre of a proper name, but no preferred or alternate term *"Bayot"* exists, nor does there exist another heading containing the same heading with another qualifier. However, the reconciliation rate actually dropped when the qualifier was omitted, due to the fact that some previously matched headings were now put into competition with others. For instance, the heading *"Models"*, which should be used as a topical subdivision under types of objects and regions of the body, had at first been matched with the same term from the Powerhouse Museum. When omitting the qualifier, headings such as *"Models (Persons)"* and *"Models (Patents)"* came into competition with the previously matched *"Models"*, thus lowering the reconciliation success. We therefore chose not to include this step within the preprocessing.

Combining the successful preprocessing steps described above, we achieved a matche of 83,117 rows (49.8%) or 61,051 records (81.1%) of the Powerhouse museum to the LCSH. Reformulating, we can state that almost half of the used terms and more than four-fifths of the collection objects have been reconciled automatically.

4.3 Reconciling the Powerhouse Museum metadata with the AAT

4.3.1 Initial reconciliation

Initially, 58,726 rows (35.2%) or 45,263 records (60.1%) were matched, if only preferred labels are considered. When alternate labels were also taken into account, match scores went up to 64,165 rows (38.4%) or 48,268 records (64.1%)

Examining the unmatched categories, several mismatches due to term qualifiers were noted. Throughout the AAT, 2,300 qualifiers are applied to 5,123 preferred and alternate terms. A limited number of qualifiers are heavily applied, such as *"species"* (302 cases) and *"wood"* (267 cases), whereas 1,663 qualifiers are applied only once. The issues regarding the use of qualifiers clearly come to the surface during the reconciliation process. For instance, the topics `300025848` and `300047753` both belong to a term with the label *"Models"*, but they have different qualifiers, *i.e.*, the former has qualifier *"(People)"* and the latter *"(Representations)"*.

The RDF conversion process follows this practice to determine the preferred and alternate labels. The initial reconciliation showed, as was the case with the LCSH, important drawbacks of qualifiers. Firstly, they may hinder automated reconciliation with a matching term even if there are no other, more specific terms present within the thesaurus. For instance, the term `300151343` *"ceramics"* has a qualifier *"objects"*, while no other term with the label *"ceramics"* exists. This is in compliance with Soergel who mentions that "all homonyms should be disambiguated with a qualifier, even if only one meaning is represented in the thesaurus" (Soergel, 1995, p. 16). However, automated reconciliation of this term failed, because the algorithm could not determine whether the Powerhouse Museum category *"ceramics"* and the AAT term *"ceramics (objects)"* refer to the same concept. Secondly, they do not help automated reconciliation if

term *"houses"* would also be plural in Spanish (*"cases"*) but singular in German (*"Haus"*) and in French (*"maison"*) (ISO25964-1, p. 27-28).

there is ambiguity. For instance, the algorithm could match the Powerhouse Museum term *"Models"* to either the AAT term *"models (people)"* or *"models (representations)"*. If both concepts had merely be labeled *"models"*, the algorithm would equally be unable to perform the match because of the two alternatives, but at least it would be able to provide two meaningful options to the user, who can then decide on a per-case basis which one is correct.

4.3.2 Enhanced reconciliation

Based on the previous insights, a second reconciliation operation was performed, this time omitting the qualifier information from the reconciliation terms. It is important to mention the positive impact on automatization of the more rigorous use of qualifiers throughout the AAT. No cases were found where both a term with and without the qualifier existed, which was responsible for the negative impact on the reconciliation success when this processing was applied with the LCSH. It is crucial to understand that the additional semantics of the qualifiers were not discarded; rather, the term text used as input for the reconciliation algorithm did not contain the qualifier. The additional semantics remain meaningful for the human who has to decide between the alternatives afterwards.

If alternate labels were also incorporated—to the extent they did not cause clashes with other preferred or alternate labels—then the success rate raised to 85,602 rows (51.3%) or 58.052 records (77.1%). Stemming for singularization was not applied, as all count nouns are expressed in plural in the AAT, which conforms to how terms are encoded in the PONT thesaurus of the Powerhouse museum. The differentiation between main headings and subdivision obviously did not apply to the AAT either. Combining the successful preprocessing steps described above, we matched more than half of the used terms and almost four-fifths of the collection objects automatically to the AAT.

5 Analysis of the reconciliation results

5.1 Summary of the different step to enhance the reconciliation process

Table 1 provides a general overview of the reconciliation rate for LCSH and AAT on a record level and allows us to understand the impact of the different methods deployed to gradually augment the success rate of the reconciliation process with both the LCSH and the AAT.

5.2 Overlap and complementarity of reconciliation with two different vocabularies

An obvious question is to what extent an overlap and a complementarity exists between the reconciliation process of the Powerhouse Museum records with both the LCSH and the AAT. Out of the total number of 167,016 rows or 75,275 records from the Powerhouse Museum collection, 59,120 rows (35.4%) or 51,524 records (68.4%) were reconciled to both the LCSH and the AAT. These figures illustrate a large overlap, which is not surprising as the LCSH was used during terminology gathering for the AAT.

However, the force of the complementarity of reconciling with more then one vocabulary is demonstrated by the number of rows and records which were only reconciled with either the LCSH or the AAT: 23,997 rows (14.3%) were only reconciled to the LCSH and did not find a match within the AAT; 26,482 rows (15.9%) were only reconciled to the AAT and did not find a match within the LCSH.

The analysis of the overlap and the complementarity can be interpreted differently when we look on the level of the unique terms, making abstraction of how many times they are used to index the collection. Out of the 4,893 unique terms used by the Powerhouse museum, 1,398 of them are matched with a heading of the LCSH (28.6%) and 1,766 are matched with a term of the AAT (36.1%). In this respect, the AAT largely outperforms the LCSH for this particular collection. Within the total amount of unique terms, 16.3% are both matched to the LCSH and the AAT, representing the overlap. The complementarity of the LCSH and the AAT on the level of unique terms is presented by the following figures: only 12.3% is exclusively matched to the LCSH (but these account for a total number of 27,191 occurrences)

	Only prefLabels	**Addition of altLabels**	**Exclusive heading**	**Omission of qualifiers**	**Stemming**
LCSH	25.5%	56.2%	78.2%	75.9%	81.1%
AAT	60.1%	64.1%	*n/a*	77.1%	*n/a*

Table 1: Overview of the different processing steps and their effect on the reconciliation rate on record level.

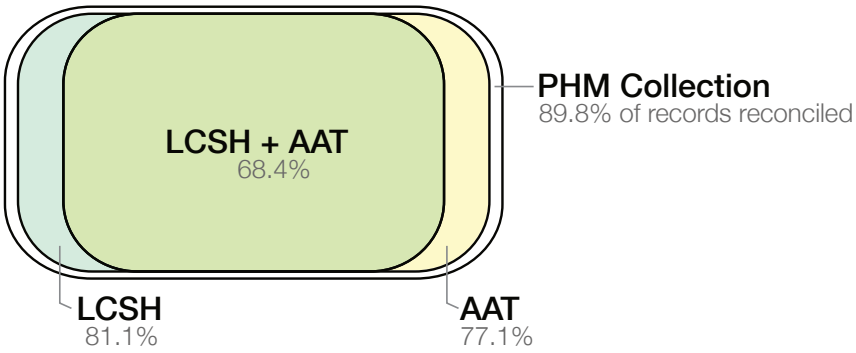

Figure 3: Almost 90% of the PHM records have been reconciled by combining the LCSH and the AAT.

and 19.8% is exclusively matched to the AAT (but only account for 29,676 occurrences). If we take into account both the LCSH and the AAT reconciliation, we can state that 109,599 out of 167,016 rows have been reconciled, or 65.6%. On the record level, this comes down to 67,579 records or 89.8% that have been automatically reconciled, as illustrated to scale in Figure 3.

5.3 Assessing the characteristics of the reconciled terms

Two questions arose about the reconciled headings: 1) how are the terms structured at the syntactic level and 2) do the reconciled headings provide a sufficient level of granularity to offer an added value in the context of information search and retrieval?

5.3.1 Syntactic structure of the reconciled terms

In order to assess the internal structure of the terms reconciled with the LCSH and the AAT, we performed a part-of-speech (POS) analysis with the help of the Natural Language Toolkit[38], a collection of Python modules for advanced text analytics, providing among other tools a probabilistic (maximum entropy) POS tagger. The tags used originate from the Penn Treebank project[39], which is the most widely established reference in the field of Natural Language Processing.

Table 2 shows the five most common structures, with figures and percentages for both the LCSH and AAT (NNS stands for plural common noun; NN for singular or mass noun; JJ for adjective and VBG for gerund, *i.e.* -ing verbal form). Terms consisting of a single plural noun (*"Flatirons"*) account for about half of all categories within both vocabularies, followed by terms formed by a plural noun modified by another noun (*"Chocolate moulds"*). Singular or mass nouns (*"Glass"*) come third for LCSH terms, but are rarer in the AAT, as could be expected from the earlier discussion of the singular/plural alternance (see section 4.3.2). More plural noun follow, modified either by an adjective (*"Acoustic guitars"*) or by a gerund (*"Copying machines"*).

In total, 43 different patterns were identified for the LCSH terms, and 38 for the AAT ones (with a large overlap between the two). These include very uncommon structures such as NN JJ NN NNS (*e.g.*, *"Gelatin dry plate negatives"*) and NN CC NN NN (*e.g.*, *"Storage and display furniture"*), which account for only one category in the dataset. Apart from the singular-noun discrepancy described above, no substantial difference was found between the LCSH and the AAT, complex terms with three words or more remaining exceptions in both vocabularies.

Two-word terms, however, when added together, represent a large number of the categories, ranging from 39.6% for the LCSH to 43.6% for the AAT. In this context, the `skos:altLabel` demonstrates its utility in the sample by linking non-preferred terms used by the Powerhouse museum such as *"Hand loom"*, *"Chocolate moulds"*, and *"Personal effects"* to the topical terms from the LCSH *"Handlooms"*, *"Chocolate molds"*, and *"Personal belongings"* respectively.

[38] `http://www.nltk.org/`
[39] `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

		LCSH		AAT	
POS tags	**Example**	#	%	#	%
NNS	*Flatirons*	655	46.9	903	51.1
NN NNS	*Chocolate moulds*	306	21.9	464	26.3
NN	*Glass*	154	11.0	59	3.3
JJ NNS	*Acoustic guitars*	80	5.7	99	5.6
VBG NNS	*Copying machines*	51	3.6	100	5.7

Table 2: Parts of speech used in the terms reconciled with the LCSH and the AAT.

5.3.2 Level of granularity and minimum level of depth

There is no formal method to categorize keywords in a deterministic manner into different levels of granularity, as this characteristic is subject to human experience and is therefore context-dependent. We should thus underline the distinction between deterministic and empirical data, and hence the type of assumptions we can draw from them. As Isabelle Boydens clearly points out, deterministic data are "characterized by the fact that there is, at any moment, a theory which makes it possible to decide whether a value (v) is correct. This is the case with algebraic data: in as much as the rules of algebra do not change over time, we can know at any time whether the result of a sum is correct. But for empirical data, which are subject to human experience, theory changes over time along with the interpretation of the values that it has made possible to determine" (Boydens, 2011, p. 113). Boydens mentions, for example, the medical domain, where theory evolves with the accumulation of experience, as witnessed, for instance, in the current research into influenza A(H1N1). Applied to the issue of keyword granularity, we could think of terms used as stop words in most domains such as "the" and "who" which could be discriminatory in the music domain when querying for "The Who".

Despite the absence of a deterministic framework to define the level of granularity, the syntactic analysis performed in the previous section allows us to focus on single- and two-word terms in order to determine if they refer to specific or generic concepts. A manual audit of a sample of reconciled categories showed that a minority of single-word terms relate to very broad and general types of objects (*e.g.*, *"Photographs"*, *"Tools"* and *"Specimens"*), whereas the majority of them deliver sufficient discriminatory value to perform interesting queries over large, heterogeneous metadata sets (*e.g.*, *"Flatirons"*, *"Carburetors"* or *"Comptometers"*, which identify highly specific object types). Most two-word terms also deliver a very precise description of the object, as illustrated by reconciled terms such as *"Babby rattles"*, *"Lawn bowls"*, *"Snuff bottles"*, *"Mustard pots"* and *"X-ray tubes"*.

However, describing terms as specific or generic remains inherently subjective. In order to have a more objective measurement of the level of specificity, we calculated the level of depth of the reconciled terms within the structure of their vocabulary. One heading can have several broader terms, which can in turn have other broader terms. One heading can therefore have different path lengths, depending on what broader term is chosen in the first place. For reasons of consistency and clarity, we decided to calculate the level of depth based upon the shortest path in the LCSH and the AAT.

We define the minimum level of depth Λ_{min} of a topic heading $t \in T$ as follows:

$$\forall t \in T : \mathit{broader}(t) = \emptyset \wedge \mathit{narrower}(t) = \emptyset \Rightarrow \quad \Lambda_{min}(t) = 0 \tag{0}$$

$$\forall t \in T : \mathit{broader}(t) = \emptyset \wedge \mathit{narrower}(t) \neq \emptyset \Rightarrow \quad \Lambda_{min}(t) = 1 \tag{1}$$

$$\forall t \in T : \min_{\lambda} (\exists b \in T : b \in \mathit{broader}(t) \wedge \Lambda_{min}(b) = \lambda) \Rightarrow \quad \Lambda_{min}(t) = \lambda + 1 \tag{2}$$

First, all headings that do not have any broader or narrower headings, and thus are not part of any hierarchy, are trivially assigned level 0 (0). Headings without broader, but with narrower headings, are assigned level 1 (1), and all other headings are assigned one level deeper than their highest direct broader heading (2).

Before analyzing the graphic which displays the level of depth of the reconciled terms within the LCSH, we need to be aware of the issues regarding the syndetic structure of the LCSH mentioned in section 1.2. The automated conversion of existing codes for cross-reference to comply with standardized thesaurus codes (USE, UF, BT, RT, SA, NT) resulted in major inconsistencies in the hierarchical relationships, as described in Dykstra (1988) and Spero (2008). The operation therefore resulted in "the appearance but not the reality of a new semantic structure" (Svenonius, 2000, p. 22). This brings us back to the conceptual differences between pre- and post-coordinated systems. As Dykstra (1988) mentions, the newly created thesaural relationship *"is a broader term of"* between the headings *"Mass media and children"* an *"Television and children"* is incorrect as this relationship only applies to the terms *"Television"* and *"Mass media"*, and has nothing to do with the term *"Children"*, which has its own thesaural relationships.

Despite the issues with the hierarchical structure of the LCSH, we believe it is useful to visualize the presence of the reconciled terms through the LCSH in Figure 4, but we should be wary of interpreting the results at face value. Almost half of the LCSH (45.6%) are positioned on level 0, and therefore do not have any broader or narrower headings, but only 9.9% of the reconciled terms belong to this group. These are components of complex subject types, such as for example *"Specimen"* which is a component of the complex subject *"Printing–Specimens"*. As these terms do not express precise concepts, it is important to know that less than 10% of the reconciled keywords belong to this level. On the other hand, Figure 4 illustrates that there is a clear match of the presence of level of depth 1, 2 and 3 values between the LCSH and the reconciled PHM terms. A high amount of reconciled terms are positioned between level 3 (9.7%), level 4 (23.8%), level 5 (4.9%), level 6 (10.6%), level 7 (6%) and level 8 (10%).

Thanks to the homogeneous groupings of terms in hierarchies, which are arranged within seven general facets, the AAT has a solid syndetic structure.We can therefore interpret the level of depth as visualized by Figure 5 in a more straightforward manner than for the LCSH. The matched terms start to appear from level 5 (1.1%) and peak in between the levels 6 (21.4%) and 8 (22.8%), after which they very gradually become less prominent starting from the level 12. On the whole, the reconciled terms find themselves 2 to 3 levels less deep when compared with the spreading of all terms throughout the AAT. This can be considered as a positive indicator of the general level of specificity of the reconciled terms.

As previously stated, we cannot claim to determine the level of granularity in an absolute, deterministic manner but Figures 4 and 5 provide at least indicators that the reconciled terms are not limited to general and broad concepts with no sufficient discriminatory value for search and retrieval.

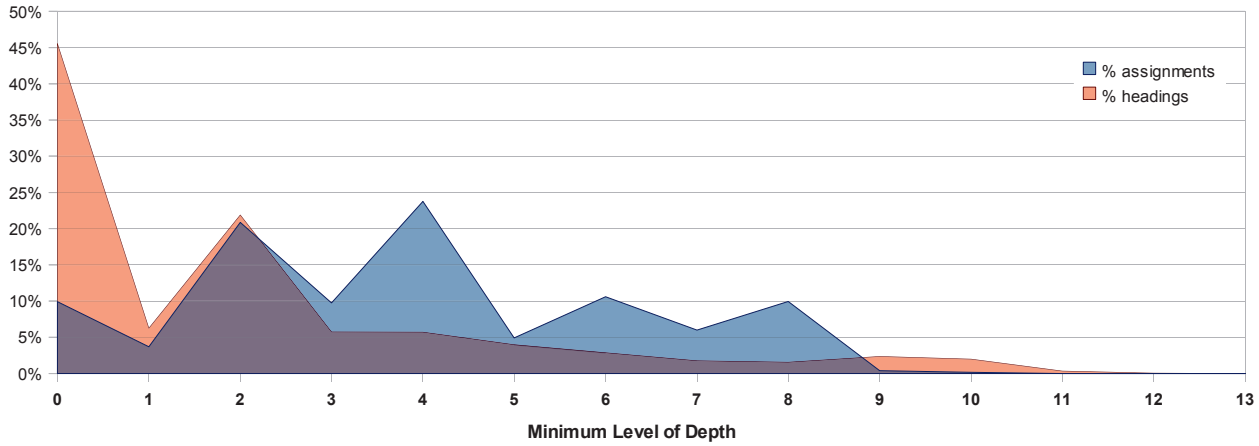

Figure 4: The level of depth assigned to the PHM collection items is generally higher than expected from the LCSH level of depth distribution.

6 Conclusions

6.1 Answers to the initial research questions

The two following questions were asked at the beginning of the paper:

- What are the possibilities to reconcile terms from a local controlled vocabulary with well-established vocabularies in an automated manner with the help of a general purpose tool for interactive data transformation?
- What are the characteristics of the reconciled terms and, more specifically, do they offer a sufficient discriminatory value for search and retrieval?

Section 4 presented in detail the results of the reconciliation process between the Powerhouse Museum records, described with the PONT, and the LCSH and the AAT. Recapitulating, we can state that about 50% of the categories and 80% of the collection objects have been reconciled with each separate vocabulary. If we aggregate the results achieved

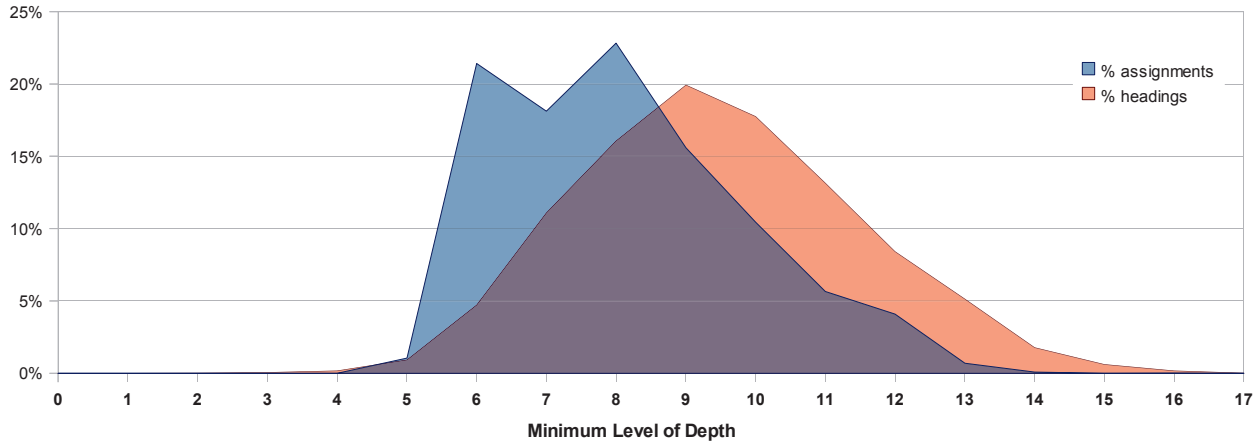

Figure 5: The level of depth assigned to the PHM collection items is slightly lower than the AAT level of depth distribution.

with the LCSH and the AAT, we reach a reconciliation rate of more than 65% of the terms. On the record level, 90% of objects have at least one category reconciled to one or both vocabularies in a fully automated way. We can therefore conclude that it is currently possible to reconcile in a satisfactory manner terms from a local thesaurus to well-established and internationally used vocabularies by using a non-expert tool.

The second research question was tackled in Section 5, in which we analyzed the characteristics of the reconciled headings and undertook the difficult process of evaluating their discriminatory value for search and retrieval. First, we had a look at the internal structure of the terms reconciled. A part-of-speech tagging analysis showed that most terms consist either of a single plural noun, or of a plural noun modified by another word (be it a noun, an adjective or a verbal form). The LCSH also exhibits a significant amount of singular nouns, which are almost absent from the AAT. Terms with three words or more exist in the collections and are occasionally reconciled with either vocabulary, but remain statistically insignificant, the most interesting group being the two-word categories which can hold a fairly specific meaning. The second analysis provided statistics regarding the level of depth of the reconciled headings. Figures 4 and 5 demonstrate that the majority of reconciled headings provide a minimum level of depth which is only slightly lower or comparable to the distribution of the level of depth for the totality of the headings in the LCSH and the totality of terms within the AAT. As the relevance and the discriminatory value of a heading is context-dependent, we cannot propose a deterministic evaluation of the possible role of the reconciled heading for search and retrieval. However, the combination of the analyses mentioned above, and especially the calculation of the minimum level of depth, make it clear that the reconciled headings are not limited to general concepts which describe an object type, such as for example "Photographs", but also comprise highly specific concepts, ranging from single-term headings such as "Flatirons" to multi-term keywords such as "Chocolate molds".

6.2 Lessons learnt from the case study

6.2.1 Recommendations for collection holders

Importance of metadata profiling and cleansing: Section 3 illustrated the importance of profiling and cleansing operations to be performed on metadata. Tools for interactive data transformation, such as Google Refine, make it easier for collection holders to perform actions such as the atomization of values in the context of field overloading, the identification of blank values and duplicate records, the application of facets and clustering to detect and filter out spelling inconsistencies.

Complementarity of different reconciliation sources: The reconciliation process with the LCSH and the AAT demonstrated both the overlap but also the complementarity of both vocabularies. Collection holders should identify the different vocabularies used within their domain, and combine both specific and general-purpose vocabularies in order to increase the success rate of the reconciliation.

Impact of preprocessing: The results of the out-of-the-box reconciliation with a general purpose tool such as Google Refine can be significantly augmented by preprocessing the in-house vocabulary and the external vocabularies used for the reconciliation. We have made the preprocessing scripts used within this paper publicly available.[30]

6.2.2 Recommendations for vocabulary managers

Use of alternate labels: The case study has demonstrated the importance of incorporating alternate labels in addition to the preferred labels. Their inclusion drastically increased the success rate of reconciliation for the LCSH (from 25.5% to 56.1% on the record level) and also boosted the matching to the AAT (from 35.2% to 38.4% on the record level).

Use of qualifiers: As prescribed by ISO25964-1, the inclusion of qualifiers should generally be avoided as much as possible. They are however essential in certain contexts, especially to distinguish homographs, yet attention should be given to a consistent manner of application. Having headings with and without qualifiers (*e.g.*, *"Models"* and *"Models (Patents)"*) in parallel within the LCSH makes the automated processing of the qualifiers more complex.

Avoid making changes to the RDF structure in which the vocabulary is offered: RDF offers a lot of possibilities and flexibility to express a certain fact in different, equivalent structures. However, a vocabulary should refrain from choosing different structures in subsequent editions, to avoid breaking existing tools.

6.2.3 Recommendations to enhance reconciliation tools

Enhancing the Google Refine RDF extension: The different preprocessing steps that we performed throughout the case study could become part of a future version of the RDF extension. Such functionalities would remove the burden for end-users, allowing them to obtain similar results without the added technical effort. By setting preferences in the interface, users could express how the RDF extension treats duplicate labels. For example, the user could give priority to preferred labels over alternate labels, and to general headings over subdivisions, which help select the correct label in the case of multiple choices. Additionally, a stemming mechanism could help with singular/plural issues.

6.3 Future work

With this paper, we hope we were successful at providing an in-depth analysis of the feasibility and benefits of semi-automated mapping methods for subject vocabularies, thereby filling a gap in the literature on metadata reconciliation.

The positive outcomes regarding our two research questions immediately lead to the next question: once the reconciliation process is mastered and understood, how to exploit the automatically created interconnections between metadata and vocabularies and achieve clear benefits for end-users and collection holders? As objects from different datasets are interlinked, they can be recommended as additional resources. Future work will focus on how browser plugins can automatically display and recommend linked resources from other collections to end-users. Collection holders can also provide a higher recall in users searches thanks to the gathering of alternate labels, and leverage the visibility of their resources since crawlers rely on links.

Recent initiatives to make collection metadata available through hosting services such as GitHub that make use of the Git revision control system, allow third parties to reuse and enrich existing metadata in an environment which makes it possible to clearly differentiate the original metadata, for which the institution takes responsibility, and the modified metadata.[40] We may investigate how version control services such as GitHub can be incorporated to easily enrich metadata with crowdsourced reconciliation efforts.

Acknowledgements

The authors would like to thank the Powerhouse Museum for making their metadata freely available and therefore allowing us to perform this case study on the basis of metadata which can be used under the CCASA license. We would also like to thank the Getty Institute for providing us with an educational license of the AAT.

The research activities described in this paper were partly funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

[40]The National Design Museum released its collection metadata as a downloadable file through Github in February 2012, using the Creative Commons Zero license. More information can be found on `http://labs.cooperhewitt.org/2012/releasing-collection-github/`.

References

M. Alistair, B. Matthews, D. Beckett, D. Brickley, M. Wilson, and N. Rogers. SKOS: A language to describe simple knowledge structures for the web. 2005. URL http://epubs.cclrc.ac.uk/bitstream/685/SKOS-XTech2005.pdf.

Julie Allinson. Openart: Open metadata for art research at the tate. *Bulletin of the American Society for Information Science and Technology*, 38(3):43–48, 2012.

James Anderson and Melissa Hofmann. A fully faceted syntax for library of congress subject headings. *Cataloging & Classification Quarterly*, 43(1):7–38, 2006.

David Bade. The perfect bibliographic record: Platonic ideal, rhetorical strategy or nonsense? *Cataloging & Classification Quarterly*, 46(1):109–133, 2009.

Tim Berners-Lee. Linked Data, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – the story so far. *International Journal On Semantic Web and Information Systems*, 5(3):1–22, 2009.

David Bodoff and Ajit Kambil. Pre-coordination + post-coordination = the case for partial coordination. Working paper is-97-14, Center for Digital Economy Research - Stern School of Business, 1997.

Isabelle Boydens. *Practical Studies in E-Government : Best Practices from Around the World*, chapter Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium, pages 113–130. Springer, 2011.

Dan Brickley and Ramanathan V. Guha. Rdf vocabulary description language 1.0: Rdf schema. Technical report, 2 2004. URL http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

Vanda Broughton. *Essential classification*. Facet Publishing, 2004.

Vanda Broughton. *Essential thesaurus construction*. Facet Publishing, 2006.

Lois Mai Chan. *Library of Congress Subject Headings. Principles and application*. Libraries Unlimited, 2005.

Martin Doerr. Semantic problems of thesaurus mapping. *Journal of Digital Information*, 1(8), 2001.

Mary Dykstra. Lc subject headings disguised as a thesaurus. *Library Journal*, pages 44–46, March 1988.

Richard Cyganiak Fadi Maali and Vassilios Peristeras. Re-using cool uris: entity reconciliation against lod hubs. In *Proceedings of the 4th Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW)*, 2011.

Philip Guo, Sean Kandel, Joseph Hellerstein, and Jeffrey Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *UIST'11, October 16-19, 2011, Santa Barbara, CA*, 2011.

Bernhard Haslhofer and Antoine Isaac. data.europeana.eu – The Europeana Linked Open Data Pilot. In *Proc. of Int. Conf. on Dublin Core and Metadata Applications*, 2011.

Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011. ISBN 9781608454303. URL http://linkeddatabook.com/.

Diane Hillmann and Jon Phipps. Application profiles: Exposing and enforcing metadata quality. In *International Conference on Dublin Core and Metadata Applications*, pages 53–62, 2007.

Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. OWL 2 Web Ontology Language Primer. W3C Recommendation, World Wide Web Consortium, October 2009. URL http://www.w3.org/TR/owl2-primer/.

Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the Pedantic Web. In *Proc. of the Linked Data on the Web (WWW2010) Workshop (LDOW 2010)*, Raleigh, North Carolina, USA, April 2010.

Antoine Isaac, Stefan Schlobach, Henk Matthezing, and Claus Zinn. Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. *Library Review*, 57(3):187 – 199, 2008. URL www.emeraldinsight.com/10.1108/00242530810865475.

ISO25964-1. ISO 25964-1 Information and documentation - thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Technical report, International Organization for Standardization, 2011.

Patrice Landry. Providing multilingual subject access through linking of subject heading languages: The macs approach. In Raffaella Bernardi and Sally Chamers, editors, *Proceedings of the workshop on advanced technologies for digital libraries*, pages 34–37, 2009.

Library of Congress Cataloging Policy and Support Office. Library of Congress Subject Headings: pre- vs post-coordination and related issues. Technical report, 2007.

Eetu Mäkelä, Osma Suominen, and Eero Hyvönen. Automatic exhibition generation based on semantic cultural content. In Lora Aroyo, Eero Hyvönen, and Jacco van Ossenbruggen, editors, *Cultural Heritage on the Semantic Web*, Bexco, Busan, Korea, November 2007. Workshop Proceedings of the 6th International Semantic Web Conference (ISWC) and 2nd Asian Semantic Web Conference (ASWC). 12. November 2007.

Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System. W3C Recommendation, 2009. URL http://www.w3.org/TR/skos-reference/.

Joachim Neubert. Bringing the "thesaurus for economics" on to the web of linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings*, volume 538, 2009. URL http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf. LDOW2009, April 20, 2009, Madrid, Spain.

Library of Congress. MADS/RDF primer, March 2011. URL http://www.loc.gov/standards/mads/rdf/.

Jack Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.

E.T. O'Neill and L.M. Chan. Fast (faceted application of subject terminology): a simplified vocabulary based on the library of congress subject headings. *IFLA Journal*, 29(4):336–442, 2003.

Juan-Antonio Pastor-Sanchez, Francisco Javier Martínez Mendez, and José Vicente Rodríguez-Muñoz. Advantages of thesauri representation with the simple knowledge organization system (SKOS) compared with other proposed alternatives for the design of a web-based thesauri management system. *Information Research*, 14(4), 2009. ISSN 1368-1613. URL http://informationr.net/ir/14-4/paper422.html.

Toni Petersen. Developing a new thesaurus for art and architecture. *Library Trends*, 38(4):644–658, 1990.

Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C recommendation, W3C, 2008. Published online on January 15th, 2008 at http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

Leo Sauermann and Richard Cyganiak. Cool URIs for the Semantic Web. World Wide Web Consortium, Note NOTE-cooluris-20081203, December 2008.

Dagobert Soergel. The art and architecture thesaurus: a critical appraisal. *Visual resources : an International Journal of Documentation*, 10(369-400), 1995.

Simon Spero. Lcsh is to thesaurus as doorbell is to mammal: Visualizing structural problems in the library of congress subject headings. In *Metadata for Semantic and Social Applications: Proceedings of the International Conference on Dublin Core and Metadata Applications*, page 203, 2008.

Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. Lcsh, skos and linked data. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2008)*, pages 25–33, 2008.

Elaine Svenonius. LCSH: Semantics, syntax and specificity. *Cataloging & Classification Quarterly*, 29(1-2):17–30, 2000. doi: 10.1300/J104v29n01_02.

Douglas Tudhope, Ceri Binding, Stuard jeffrey, Keith May, and Andreas Vlachidis. A stellar role for knowledge organization systems in digital archaeology. *Bulletin of the American Society for Information Science and Technology*, 37(4):15–18, April/May 2011.

Lourens van der Meij, Antoine Isaac, and Claus Zinn. A web-based repository service for vocabularies and alignments in the cultural heritage domain. In *Proceedings of the 7th European Semantic Web Conference (ESWC)*, volume 6088, pages 394–409, 2010.

Marieke van Erp, Johan Oomen, Roxane Segers, Chiel van den Akker, Lora Aroyo, Geertje Jacobs, Susan Legêne, Lourens van der Meij, Jacco van Ossenbruggen, and Guus Schreiber. Automatic heritage metadata enrichment with historic events. In J. Trant and D. Bearman, editors, *Museums and the Web 2011: Proceedings*. Archives & Museum Informatics, Toronto, 2011.

Seth van Hooland, Seth Kaufman, and Yves Bontemps. Answering the call for more accountability: applying data-profiling to museum metadata. In *International conference on Dublin Core and metadata applications, 22- 26 september 2008*, pages 93–103, 2008.

Seth van Hooland, Eva Mendez, and Françoise Vandooren. Opportunities and risks for libraries in applying for european funding. *The Electronic Library*, 29(1):90–104, 2010.

Seth van Hooland, Eva Mendez, and Isabelle Boydens. Between commodification and sense-making. on the double-sided effect of user-generated metadata within the cultural heritage sector. *Library Trends*, 59(4):707–720, 2011.

Leonard Will. The iso 25964 data model for the structure of an information retrieval thesaurus. *Bulletin of the American Society for Information Science and Technology*, 38(4):48–51, April/May 2012.

Kwan Yi and Lois Mai Chan. Linking folksonomy to library of congress subject headings: an exploratory study. *Journal of Documentation*, 65(6):872–900, 2009. ISSN 0022-0418. doi: 10.1108/00220410910998906. URL `http://www.emeraldinsight.com/journals.htm?articleid=1823651&show=html`.