Building your own LOCKSS-based distributed preservation network

Workshop - ILIDE conference 2018

Anthony Leroy
Université libre de Bruxelles

Agenda

Motivation

LOCKSS main concepts

SAFE LOCKSS network : a Case Study

Build your own LOCKSS Network

Discussions

Agenda

Motivation

LOCKSS main concepts

SAFE LOCKSS network : a Case Study

Build your own LOCKSS Network

Discussions

LEARNING RESOURCES

University Libraries have two main **missions**

LEARNING RESOURCES

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

LEARNING RESOURCES

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

For analog objects, guaranteeing access and preservation is relatively simple

For analog objects, guaranteeing access and preservation is relatively simple

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

In the digital era, those missions are compromised:
- we lost control on some digital objects (access via subscription)
- the vulnerability of digital objects

University Libraries have two main **missions**

To guarantee access to objects
selected by curators

To preserve those objects
especially our own production

Post-cancellation access to electronic publications is often considered as a big preservation problem for libraries

Post-cancellation access to digital publications is **a big issue** for libraries

STANFORD UNIVERSITY

UNIVERSITEIT GENT

Universität Bielefeld

Post-cancellation access to digital publications is **a big issue** for libraries

STANFORD UNIVERSITY

?

?

UNIVERSITEIT GENT

?

Universität Bielefeld

?

Hopefully, three efficient post cancellation access solutions are available

CLOCKSS

+240 publishers

ELSEVIER · Springer · WILEY · npg · SAGE · OXFORD UNIVERSITY PRESS · Taylor & Francis Taylor & Francis Group

+7400 titles

dark archive

$500-$15000

Hopefully, three efficient post cancellation access solutions are available

CLOCKSS

PORTICO

+240 publishers

ELSEVIER · Springer · WILEY · npg · SAGE · OXFORD UNIVERSITY PRESS · Taylor & Francis Group

+7400 titles

dark archive

$500-$15000

+510 publishers

ELSEVIER · WILEY · npg · Springer · Taylor & Francis Group · SAGE · CAMBRIDGE UNIVERSITY PRESS · Hindawi · IEEE · OXFORD UNIVERSITY PRESS

+28000 titles

dark archive

$1500-$25000

Hopefully, three efficient post cancellation access solutions are available

CLOCKSS

PORTICO

LOCKSS

+240 publishers

ELSEVIER, Springer, WILEY, npg, SAGE, OXFORD UNIVERSITY PRESS, Taylor & Francis Taylor & Francis Group

+7400 titles

dark archive

$500-$15000

+510 publishers

ELSEVIER, WILEY, npg, Springer, Taylor & Francis Taylor & Francis Group, SAGE, CAMBRIDGE UNIVERSITY PRESS, Hindawi, IEEE, OXFORD UNIVERSITY PRESS

+28000 titles

dark archive

$1500-$25000

+530 publishers

Springer, Hindawi, SAGE, Emerald, OXFORD UNIVERSITY PRESS, INDER SCIENCE, CAMBRIDGE UNIVERSITY PRESS, Taylor & Francis Taylor & Francis Group

+12900 titles

light archive

$1080-$10800

LEARNING RESOURCES

Join the LOCKSS Alliance, support CLOCKSS and your digital preservation problem is solved !

LEARNING RESOURCES

University Libraries have two main **missions**

To guarantee access to objects selected by curators

To preserve those objects
especially our own production

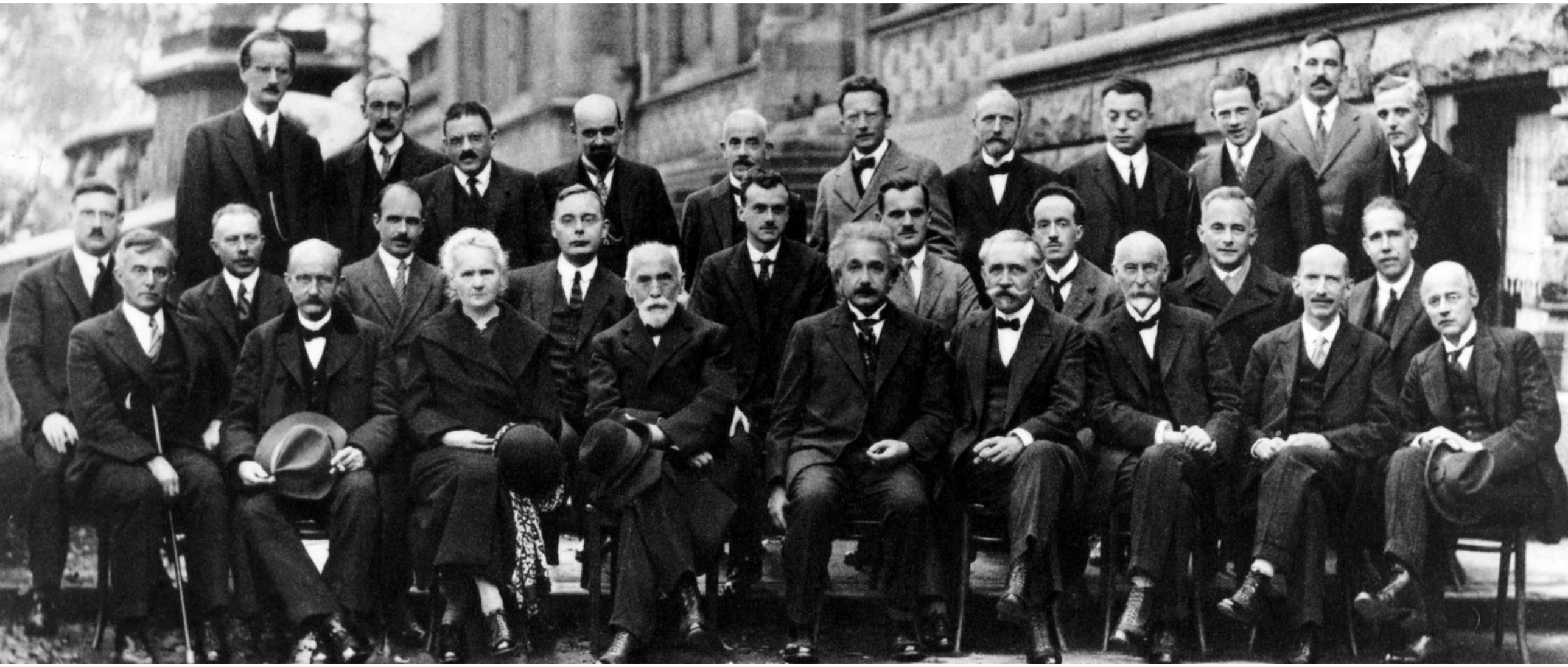
University Libraries have two main **missions**

To guarantee access to objects selected by curators

To preserve those objects
especially our own production

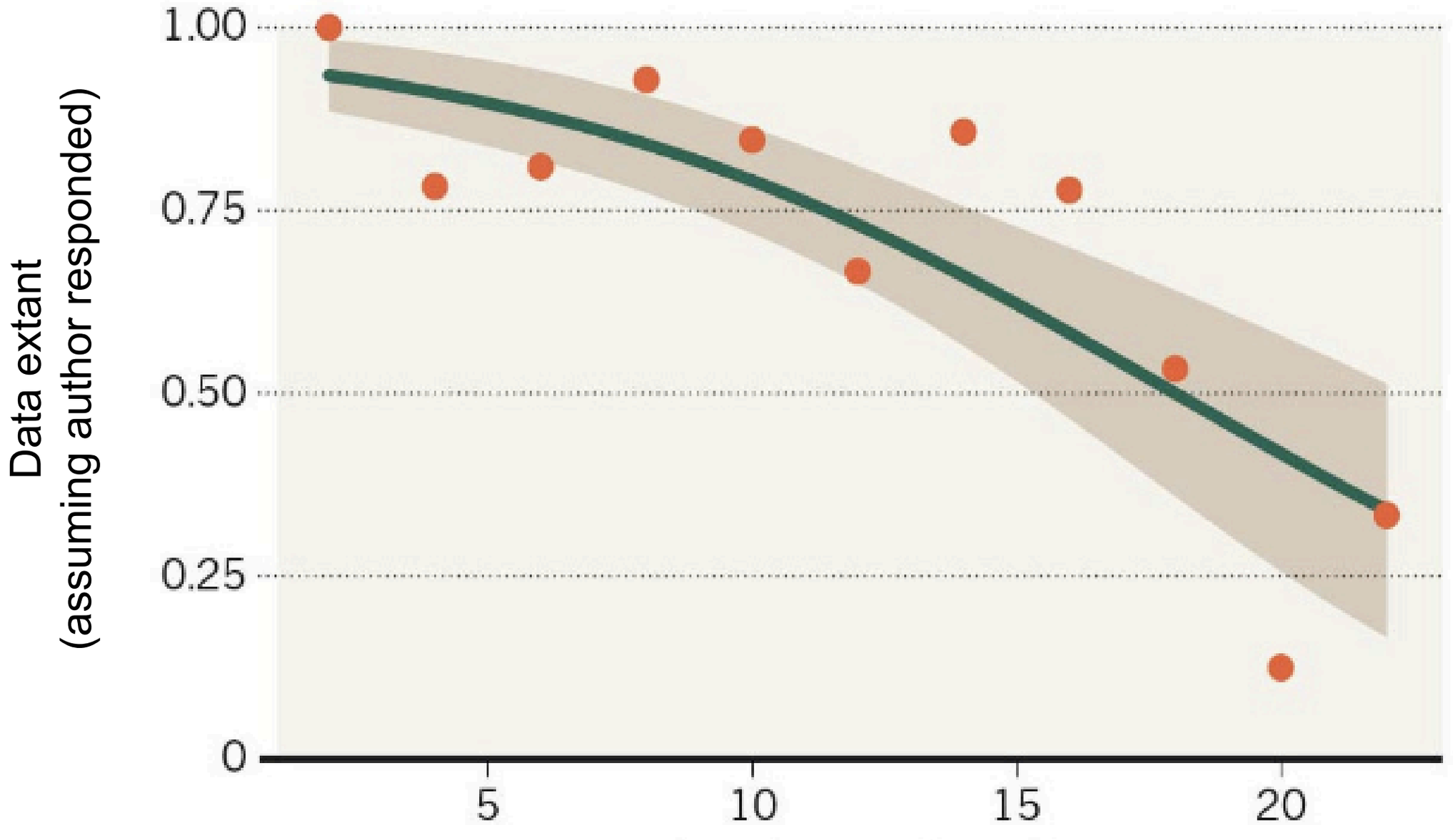
Still, there is one crucial problem to solve:
How are we going to preserve our own digital objects collections?

What will happen to the current research outcome in 10 yrs? 100 yrs?



Attendees of the 5th Solvay Congrès, October 1927, Institut international de physique Solvay, Brussels

Massive loss of research data in the last 25 years



80% of data in zoology linked to publications in the nineties are definitely lost

Vines Et Al., The Availability Of Research Data Declines Rapidly With Article Age, Current Biology (2014), Http://Dx.Doi.Org/10.1016/J.Cub.2013.11.014

Scientific knowledge results from a complex iterative process

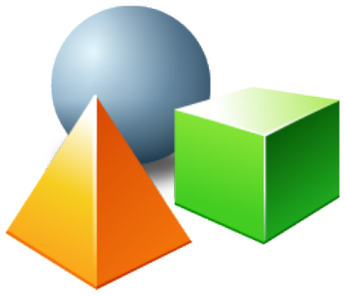
[Encyclopedic manuscript containing allegorical and medical drawings], Library of Congress, Rosenwald 4, Bl. 5r

"If I have seen further it is by standing on the shoulders of Giants"

Isaac Newton, 1676

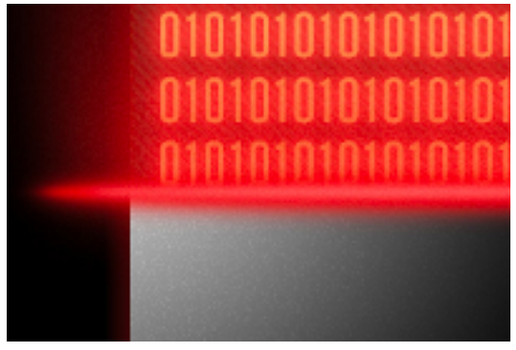
As a digital library, our core business is to develop the software infrastructure to describe, submit, disseminate and preserve digital objects

Dissemination



Digital object (digitized)

Description



Submission or Digitization

Preservation

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

Not to be confused with :

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

Not to be confused with :

- **storage:** recording data on a physical medium

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

Not to be confused with :

- **storage:** recording data on a physical medium
- **backup :** replicating data in order to restore them quickly in case of loss

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

Not to be confused with :

- **storage:** recording data on a physical medium
- **backup :** replicating data in order to restore them quickly in case of loss
- **archiving in the ordinary IT meaning**: moving less frequently used data on cheaper storage media (typically magnetic tapes)

What exactly is digital preservation ?

Strategies and processes to **protect** against the **threats** endangering digital objects of interest with the aim of (**re**)**using** them in the future on the **very long term**.

Not to be confused with :

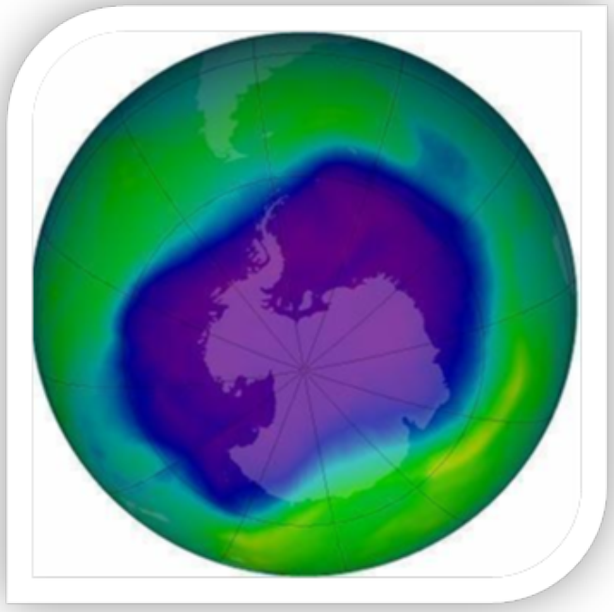
- **storage:** recording data on a physical medium
- **backup :** replicating data in order to restore them quickly in case of loss
- **archiving in the ordinary IT meaning**: moving less frequently used data on cheaper storage media (typically magnetic tapes)

An effective digital preservation solution should thus be based on a threat-model.

Two **levels** of preservation

bit-level preservation

logical-level preservation



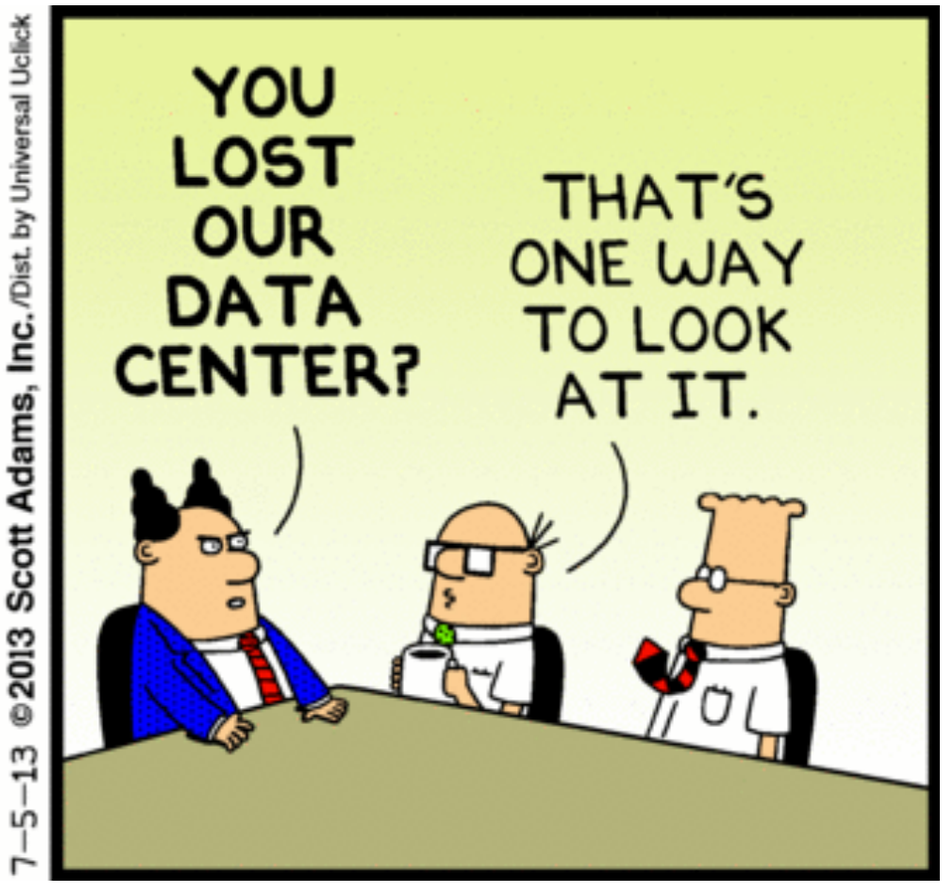
Two **levels** of preservation

bit-level preservation

logical-level preservation

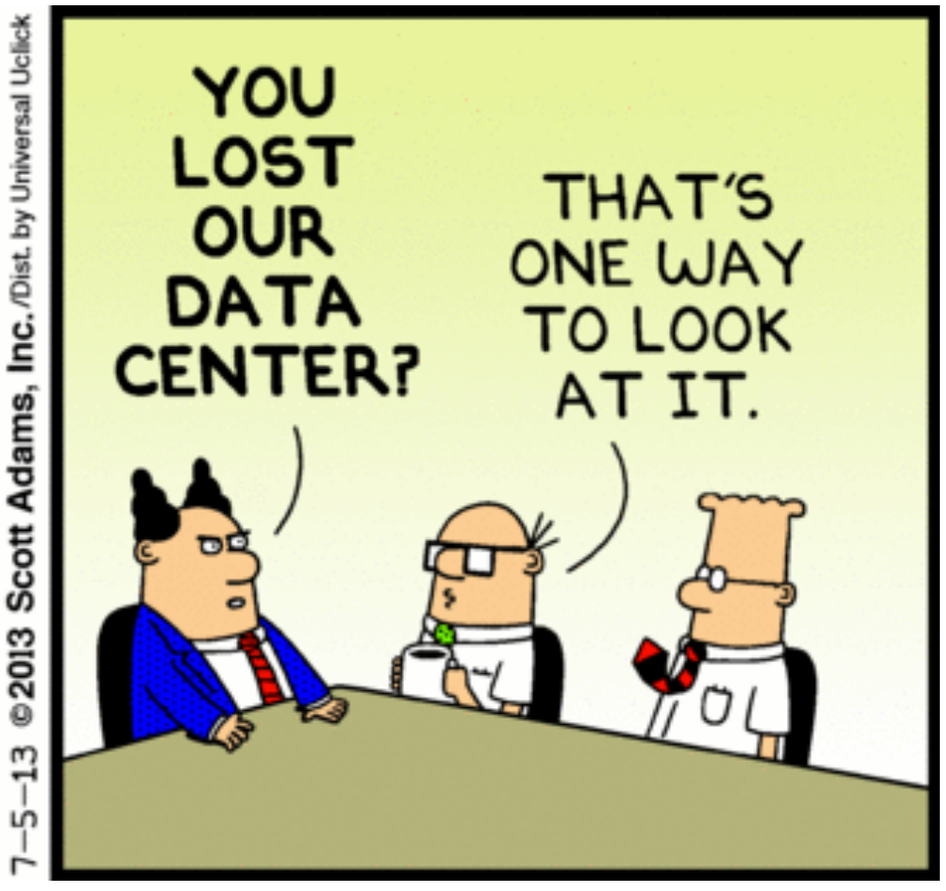
What threats are endangering our digital resources and what can we do about it ?

What threats are endangering our digital resources and what can we do about it ?



THREAT

What threats are endangering our digital resources and what can we do about it ?



THREAT

SOLUTION

Dilbert, Scott Adams

1. Natural and man-made disasters

https://flic.kr/p/9MsNf9

1. Natural and man-made disasters

https://flic.kr/p/9MsNf9

Very difficult risk to evaluate… low-probability with major impact events

It happens…

GOOGLE

Google suffers data loss as data center gets hit by lightning 4 times

BY NATALIE SHOEMAKER 08.21.2015 :: 9:30AM EST

Fire Causes Partial UniSuper Data Center Outage in Port Melbourne

As Irma Heads for Florida, One Miami Data Center is Especially Critical

While an outage is far from given, its effects would ripple far beyond cities in the hurricane's path.

Michigan County Offline After Data Center Fire

IT ser
after
count
west
have

Rich Mi

Toronto Flooding KOs Data Center Cooling Systems

A massive rainstorm caused widespread flooding and power outages Monday night in Toronto, which created challenges for some tenants at the city's largest data center hub. 151 Front Street maintained power, but experienced problems with cooling systems.

Kaseya, ConnectWise Pick Up Pieces After Irma Nails Florida HQs

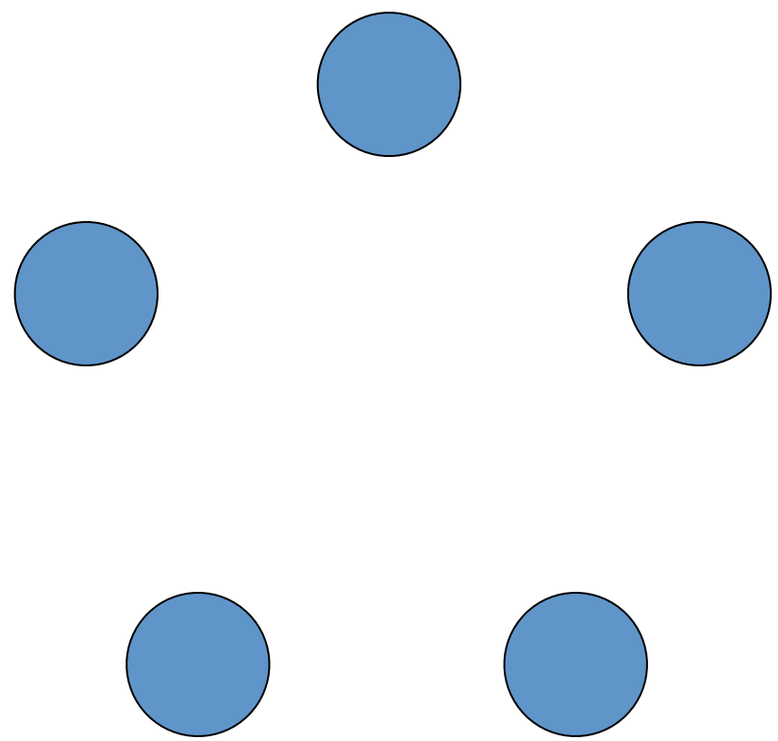
Both companies supporting customers, but operations not expected to return to normal for days

Aldrin Brown | Sep 13, 2017

http://www.datacenterknowledge.com/

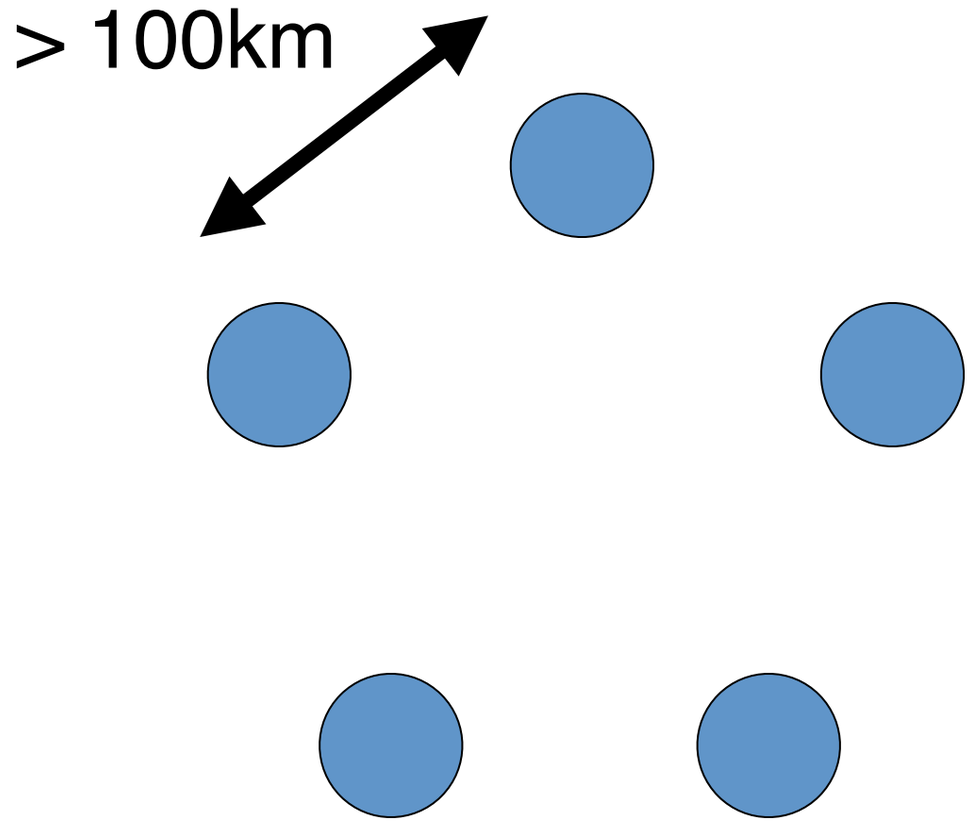
1. Natural and man-made disasters

→ Geo-replication



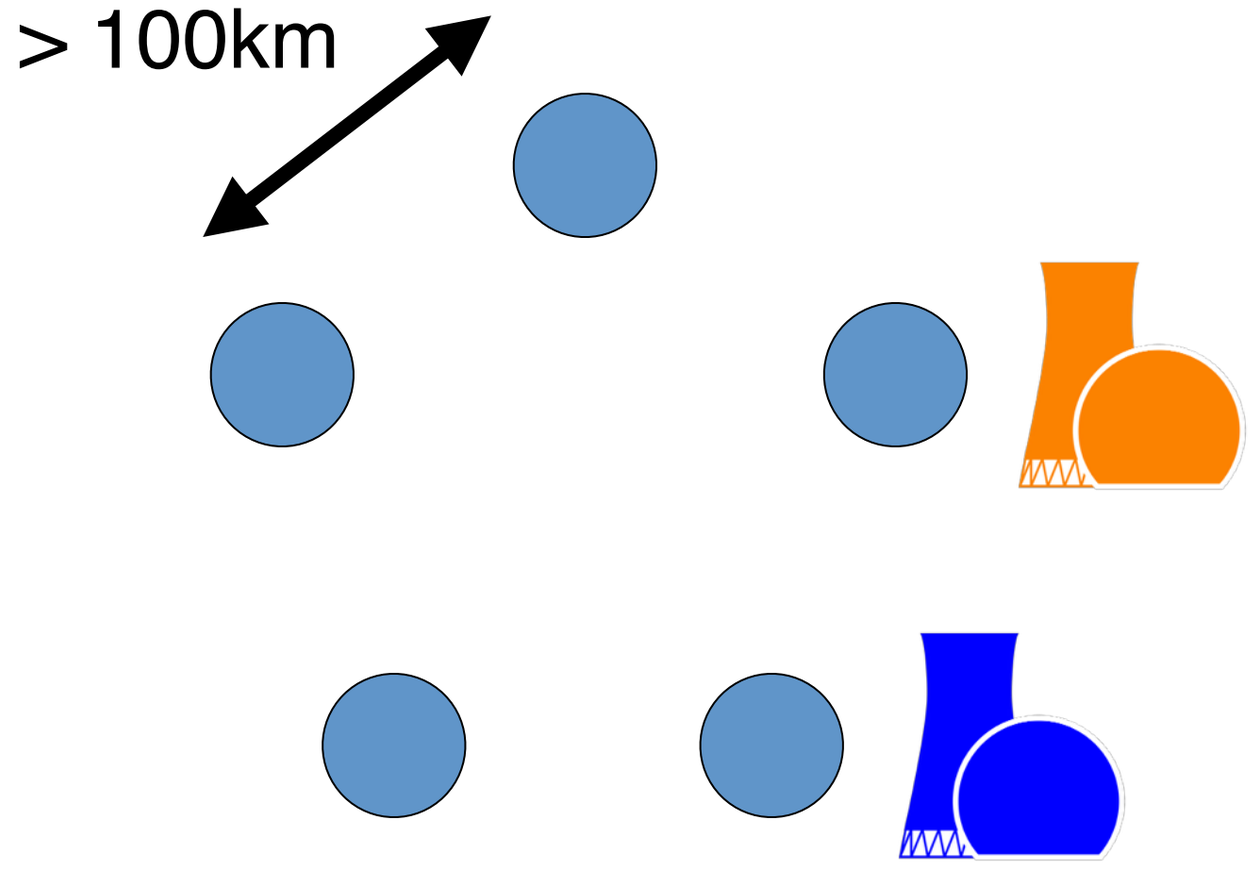
1. Natural and man-made disasters

→ Geo-replication



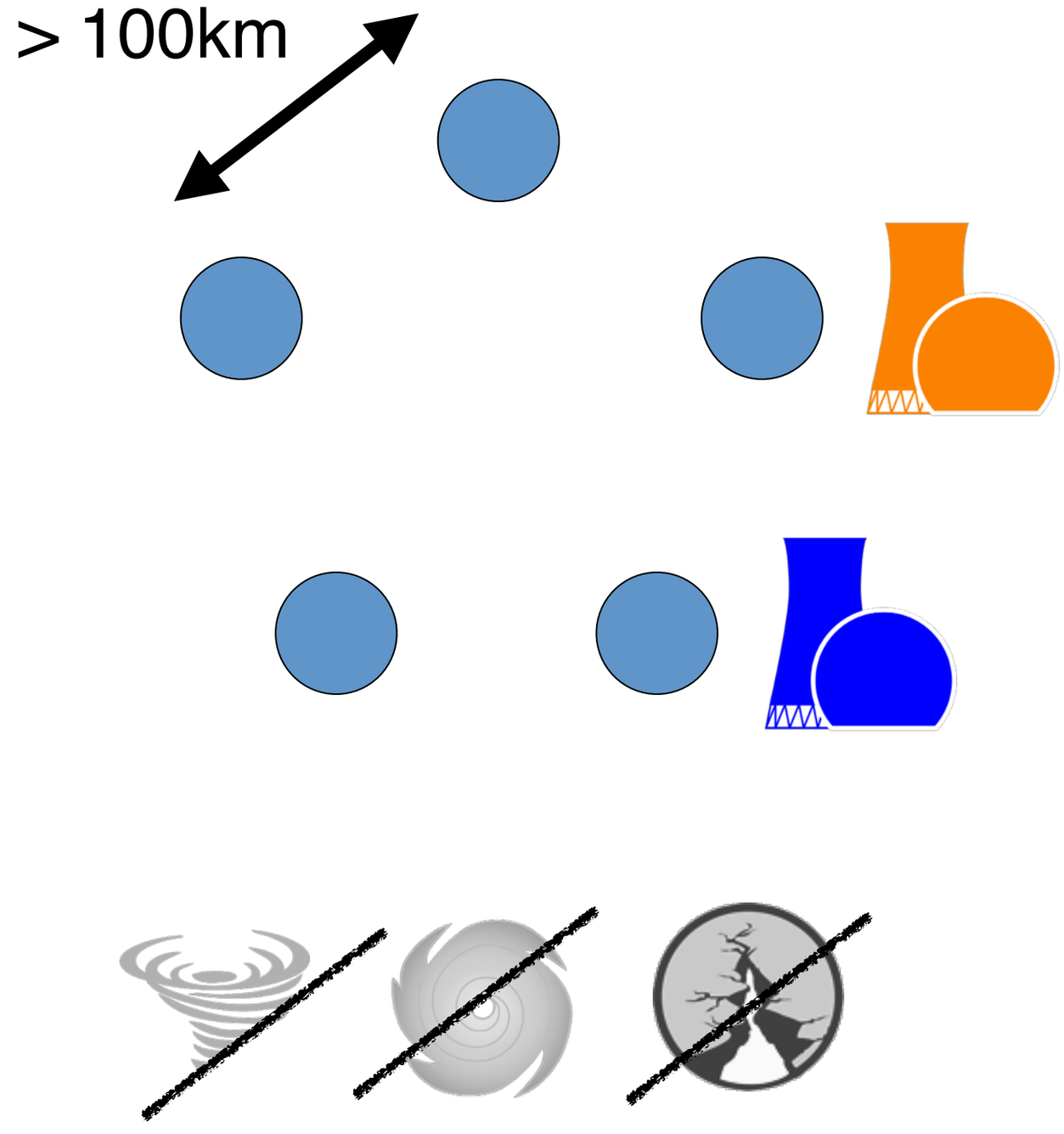
1. Natural and man-made disasters

→ Geo-replication

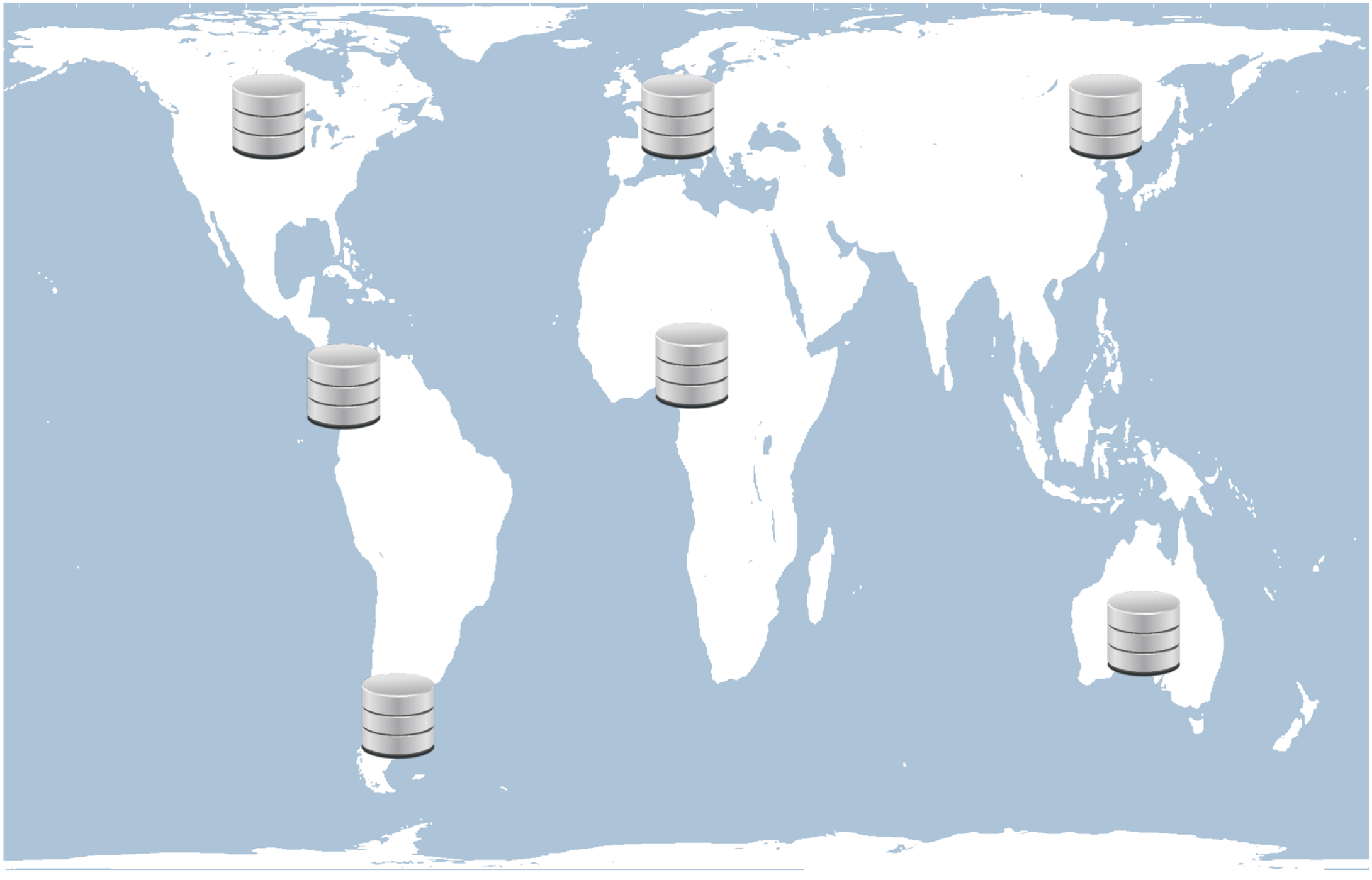


1. Natural and man-made disasters

→ Geo-replication

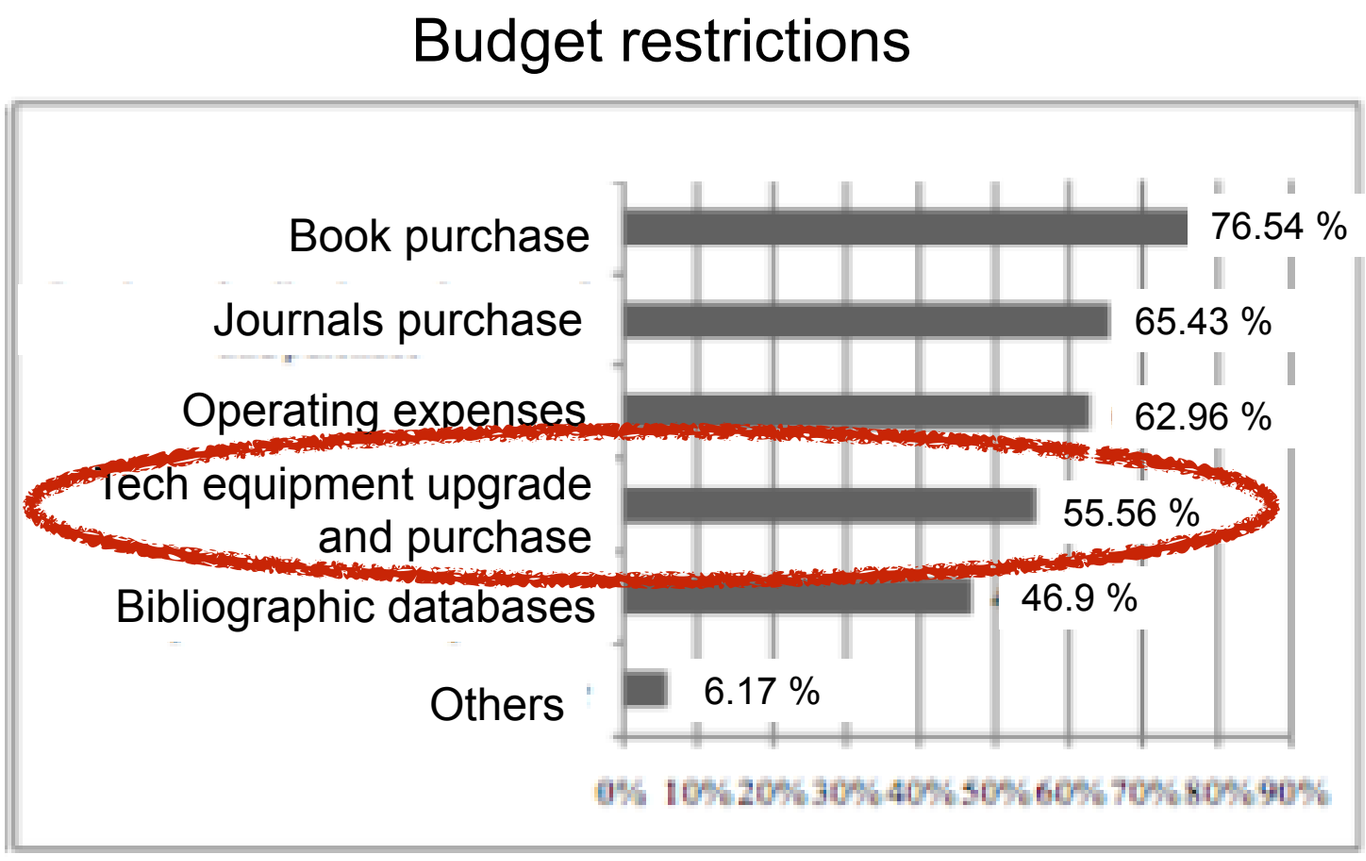


The ideal solution : geo-replication of copies



2. Economic risk

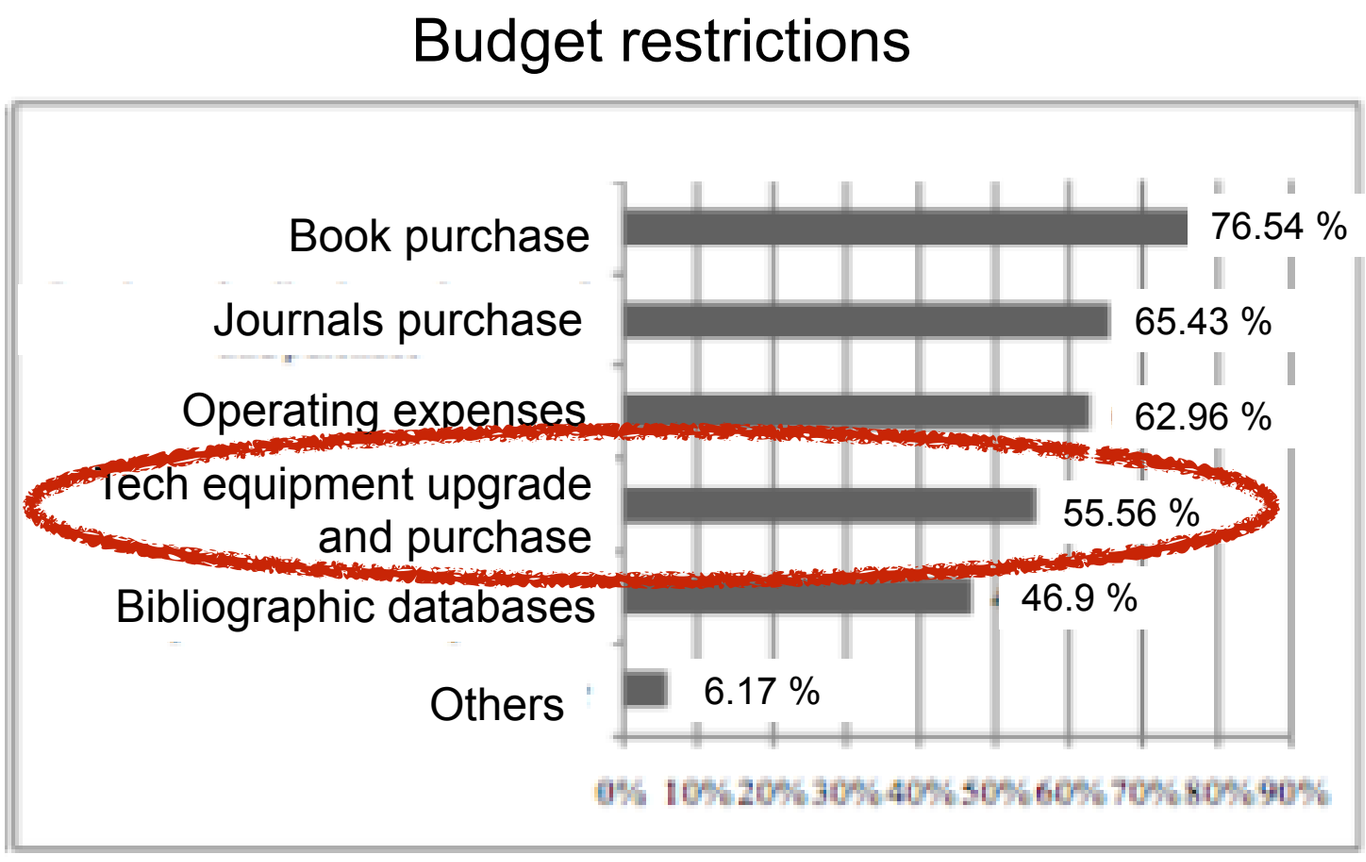
In times of economic crisis, budget restrictions first and foremost concern activities that do have low impact on the short term future



G. Giannakopoulos et al., Libraries in Crisis: A Glimpse over Greece and Cyprus, Procedia, Vol.147, 2014, https://doi.org/10.1016/j.sbspro.2014.07.121.

2. Economic risk

In times of economic crisis, budget restrictions first and foremost concern activities that do have low impact on the short term future



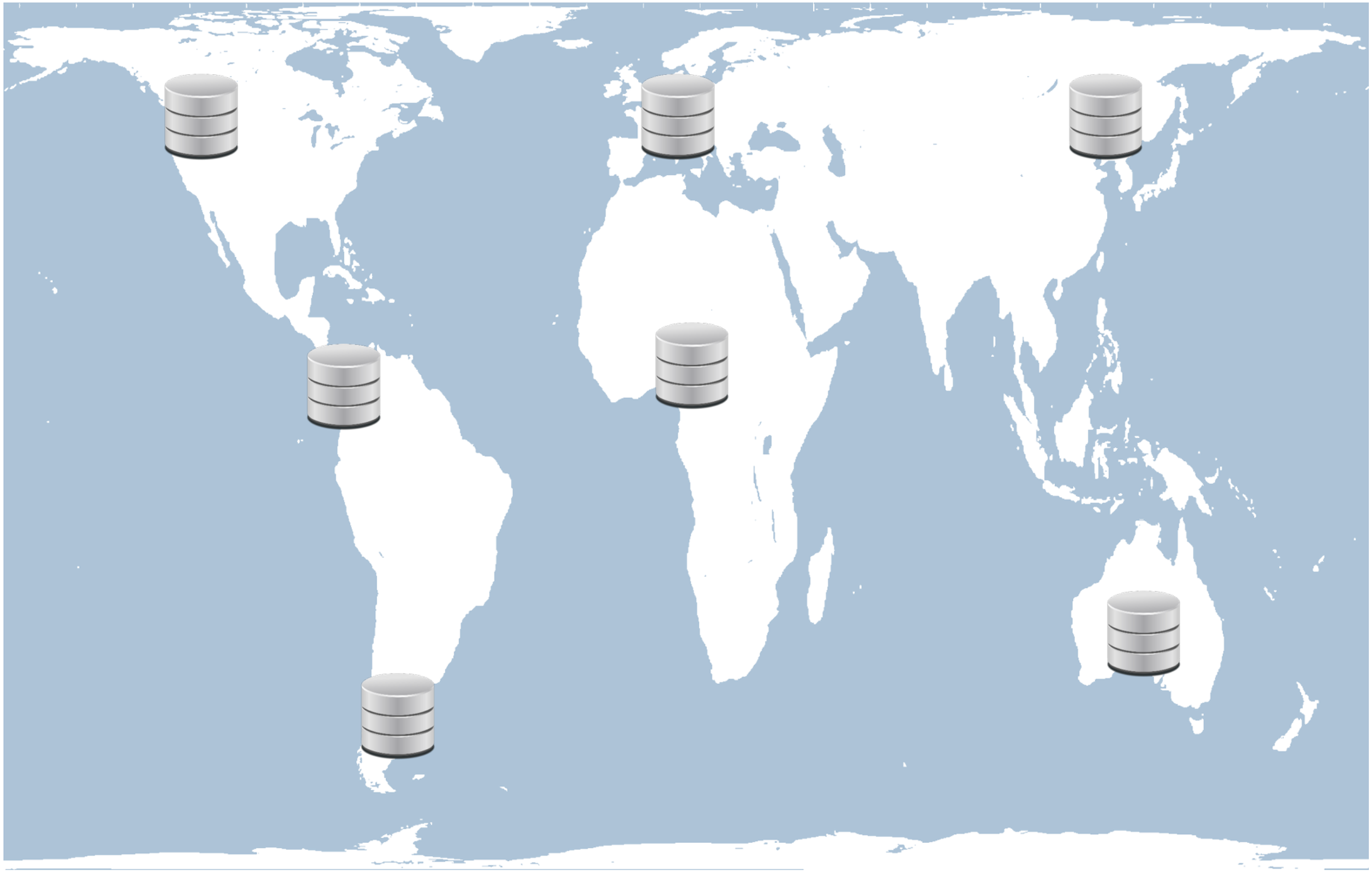
G. Giannakopoulos et al., Libraries in Crisis: A Glimpse over Greece and Cyprus, Procedia, Vol.147, 2014, https://doi.org/10.1016/j.sbspro.2014.07.121.

If investments in the infrastructure maintenance cannot be done for just 5 years, the risk of data loss considerably increases.

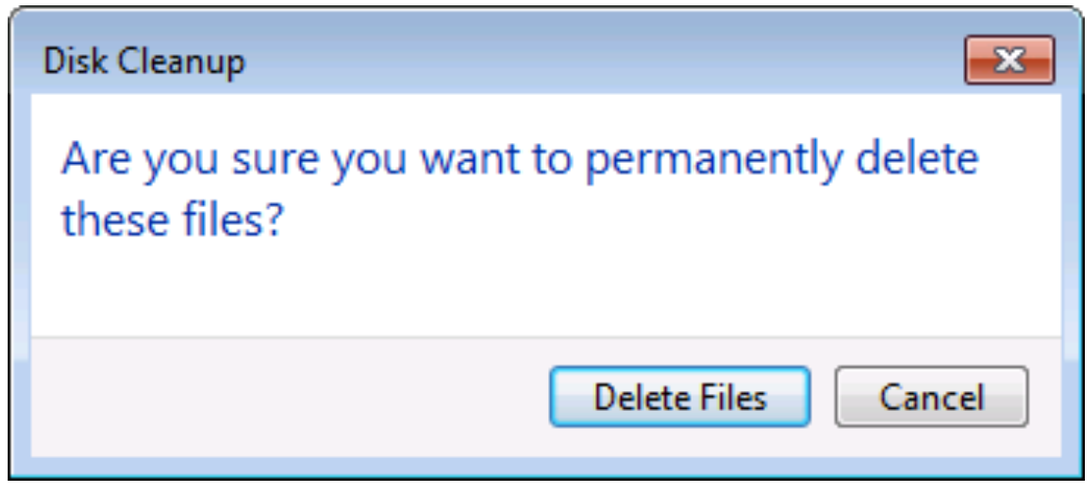
LEARNING RESOURCES

2. Economic risk

→ Cost control



3. Human errors



It happens…

Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data

Henry Blodget

Apr. 28, 2011, 7:10 AM 112,400

AWS Outage that Broke the Internet Caused by Mistyped Command

Amazon says Tuesday's mayhem resulted from mistake during a routine debugging exercise.

Yevgeniy Sverdlik | Mar 02, 2017

British Airways: Engineer Wrongly Disconnected Data Center Power Supply

How a single tech could cause so much damage, and why the backup system failed remain open questions.

Bloomberg | Jun 06, 2017

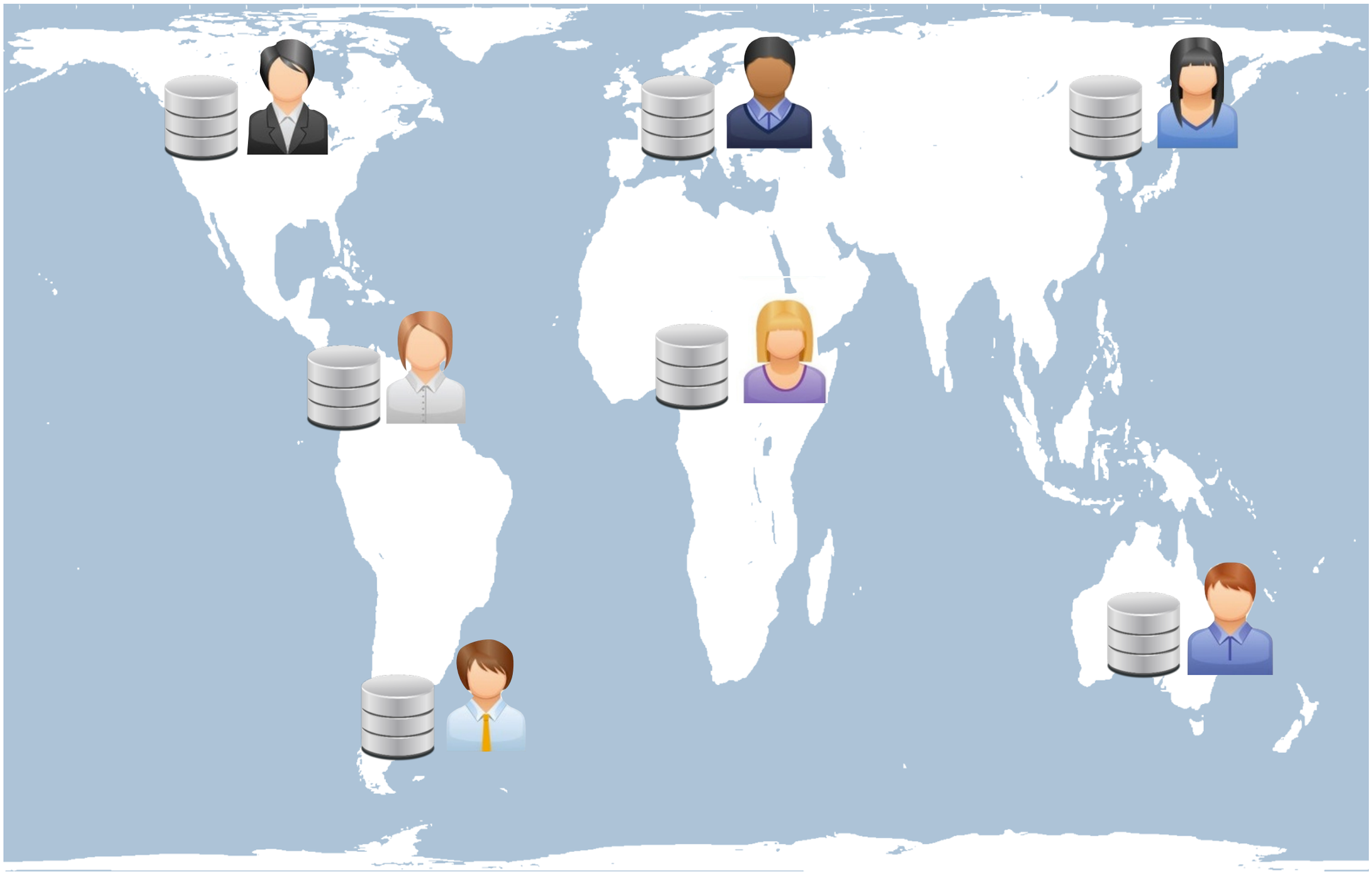
Human Error Cited in Hosting.com Outage

Hosting.com said human error was responsible for a data center power outage that left more than 1,100 customers without service. The downtime occurred as the company was conducting preventive maintenance on a UPS system in the company's data center in Newark, Del.

Rich Miller | Jul 28, 2012

3. Human errors

→ Individual technical management



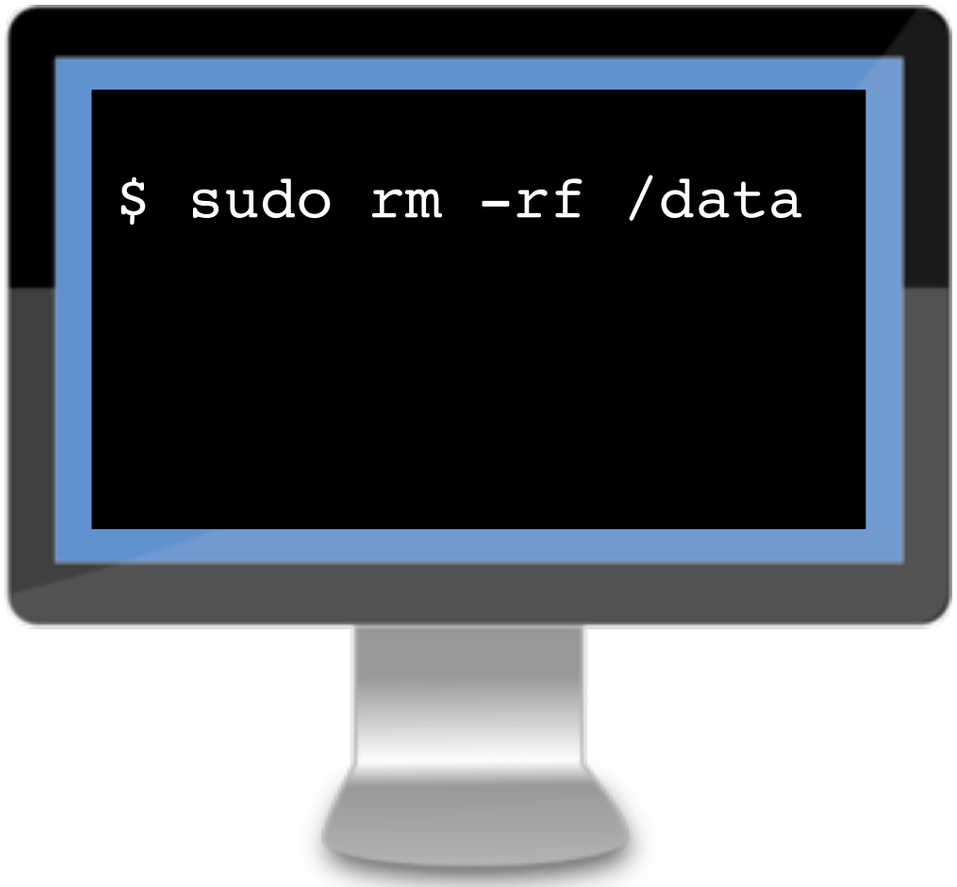
4. Organizational, politic and ideologic risk

A digital autodafé is much more efficient than its analog equivalent

In the 15th century

A digital autodafé is much more efficient than its analog equivalent

In the 15th century



Today

The scientific heritage is not spared

Vanishing Canada: Why we're all losers in Ottawa's war on data

Records deleted, burned, tossed in Dumpsters. A Maclean's investigation on the crisis in government data

Anne Kingston

September 18, 2015

WILDLIFE WATCH

U.S. Animal Abuse Records Deleted—What We Stand to Lose

By hiding online records of welfare violations, U.S. agency robs journalists, investigators, and the public of timely information—and takes pressure off abusers.

BUSINESS INSIDER UK

POLITICS

The Trump administration has told the EPA to remove its climate-change data from its website

Rafi Letzter

Jan. 25, 2017, 3:55 PM 1,781

Sections

The Washington Post

Energy and Environment

Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump

By Brady Dennis December 13, 2016

Most Read

LEARNING RESOURCES

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

LEARNING RESOURCES

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

- No control on the archiving technical policy

LEARNING RESOURCES

We want to be “active players of preservation”, not “passive clients of third-party preservation services“ [Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

- No control on the archiving technical policy
- Is migration to another provider possible?

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

No more technical problems...

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

- No control on the archiving technical policy
- Is migration to another provider possible?

at the cost of many legal and economical issues

LEARNING RESOURCES

We want to be “active players of preservation”,
not “passive clients of third-party preservation services“
[Skinner11]

Our only guarantee of preservation would be the SLA

Legal issues:

- What if the service provider goes bankrupt?
- What if data gets lost? Can we claim for damages?

Technical issues:

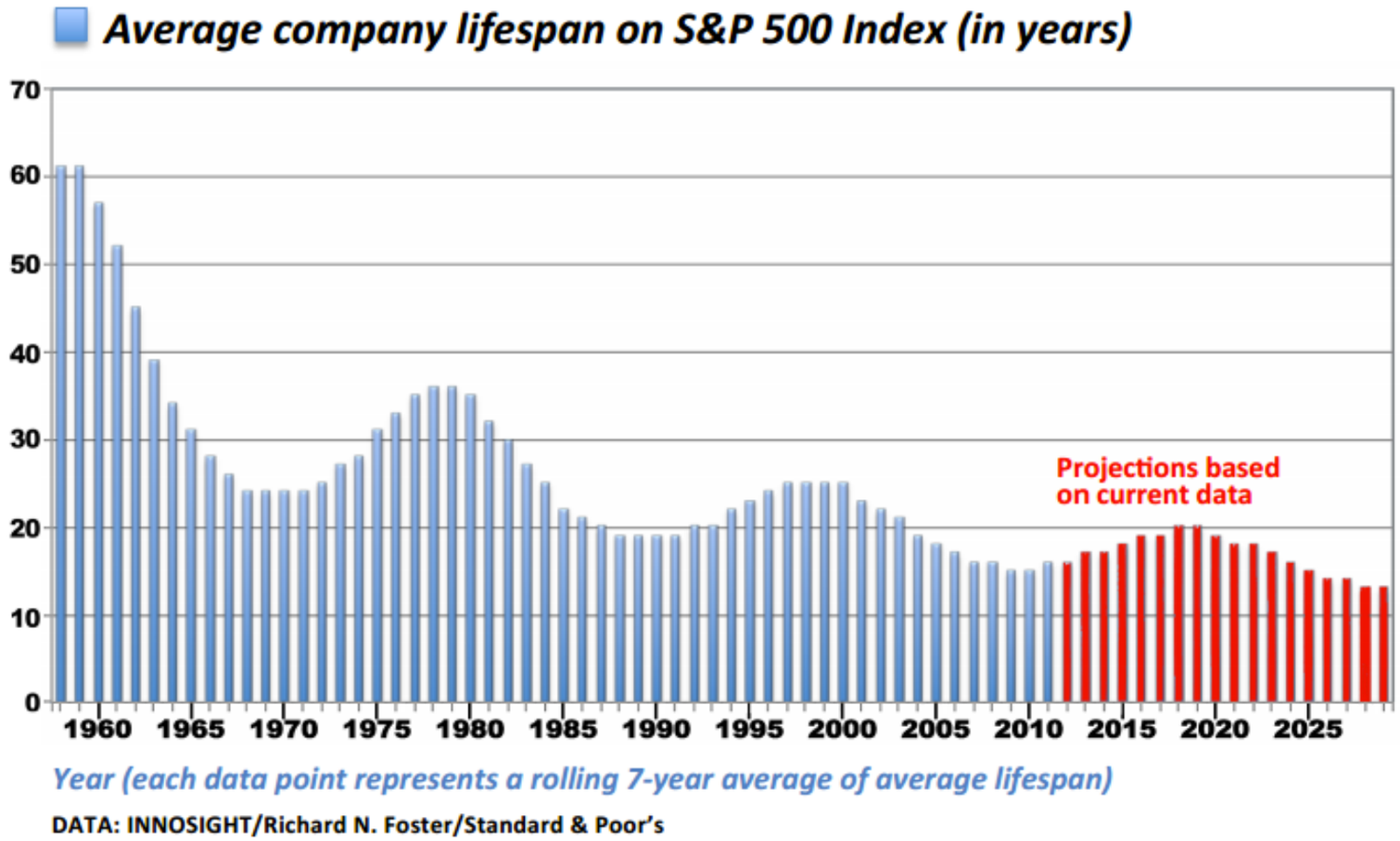
- No control on the archiving technical policy
- Is migration to another provider possible?

No more technical problems...

at the cost of many legal and economical issues

Control is key in digital preservation

Universities are amongst the oldest institutions around



What organization is better qualified to preserve data on the long term?

Cloud storage : yet another oligopole …

Cloud storage : yet another oligopole …

amazon web services

Microsoft Azure

IBM Cloud

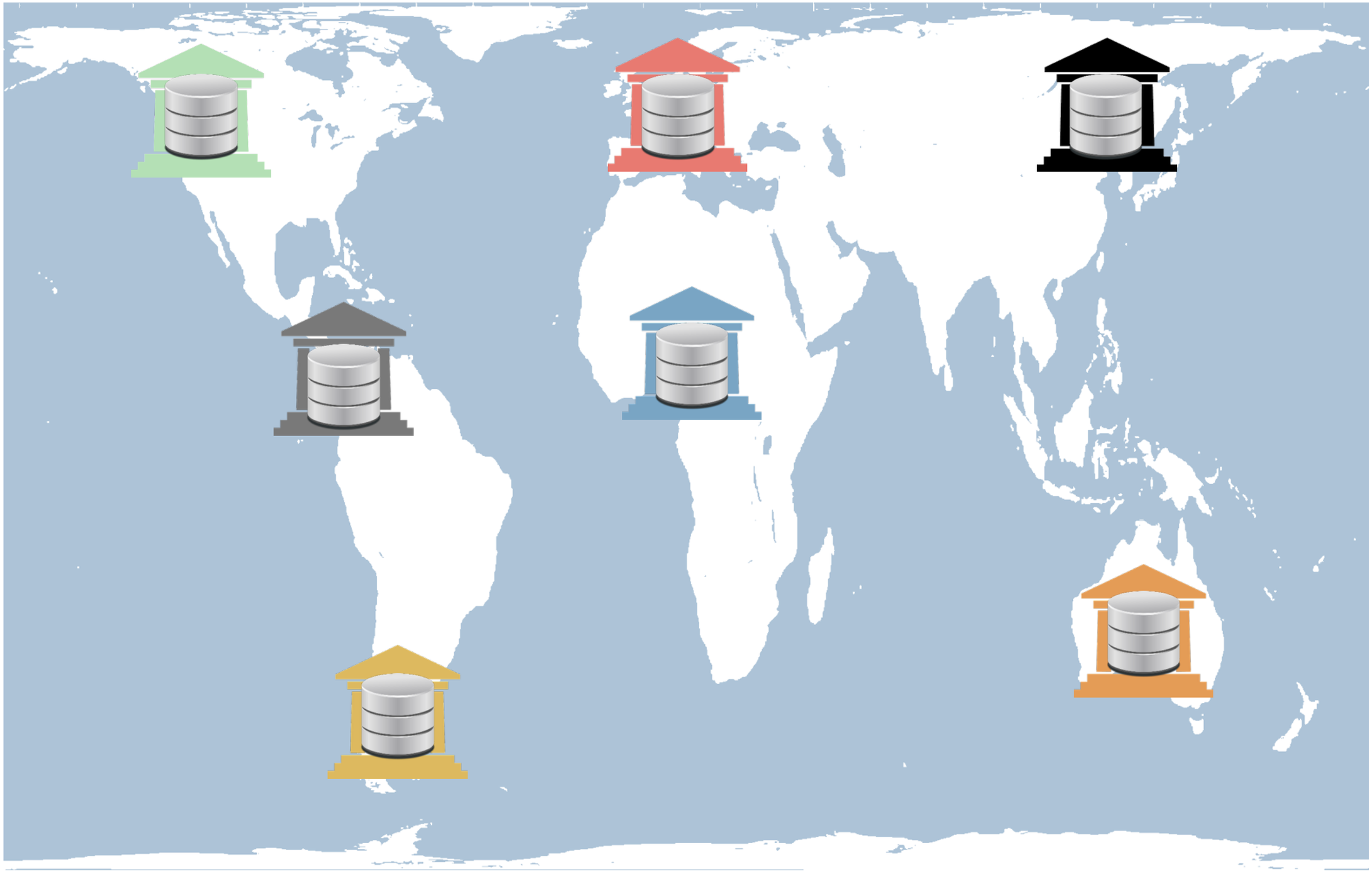
Google Cloud Storage

IaaS/PaaS market (Source: Jefferies)

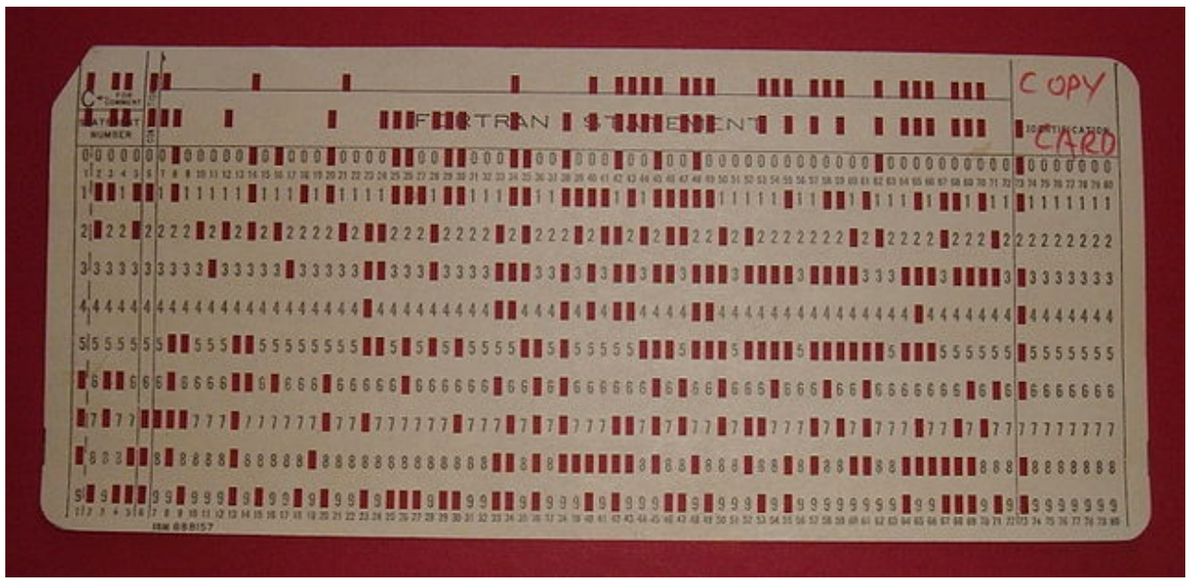
Cloud vendor	Annualized revenue	% of market	Year-over-year growth
Amazon Web Services	$18.34 billion	51%	42%
Microsoft Azure	$6.17 billion	17%	89%
IBM Cloud	$4.03 billion	11%	22%
Google Cloud Platform	$2.05 billion	6%	125%
Alibaba Cloud	$1.79 billion	5%	92%
Salesforce	$1.78 billion	5%	31%
Oracle Cloud	$1.59 billion	4%	20%
Subtotal	$35.75 billion	86%	54%
Total Gartner estimate	$41.79 billion	100%	33%

4. Organizational risk

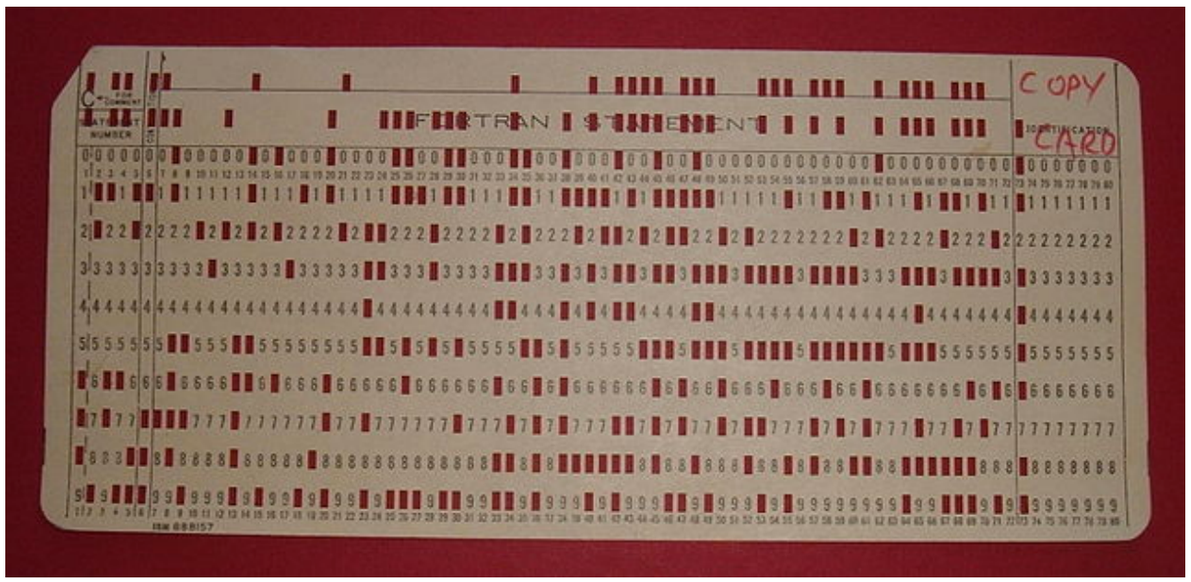
→ Administrative independence



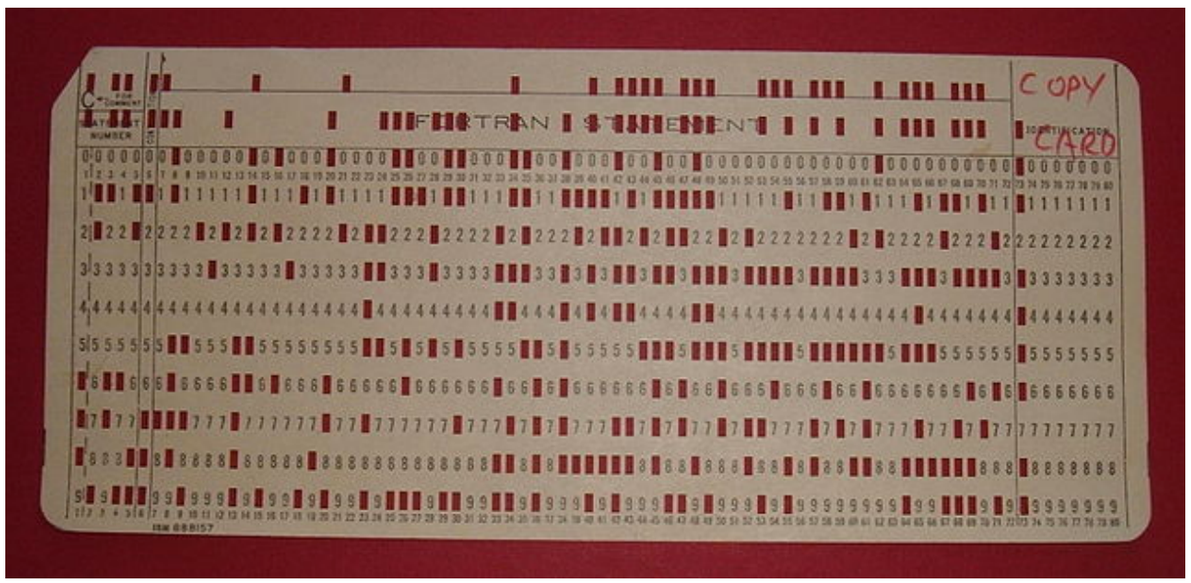
5. Hardware obsolescence



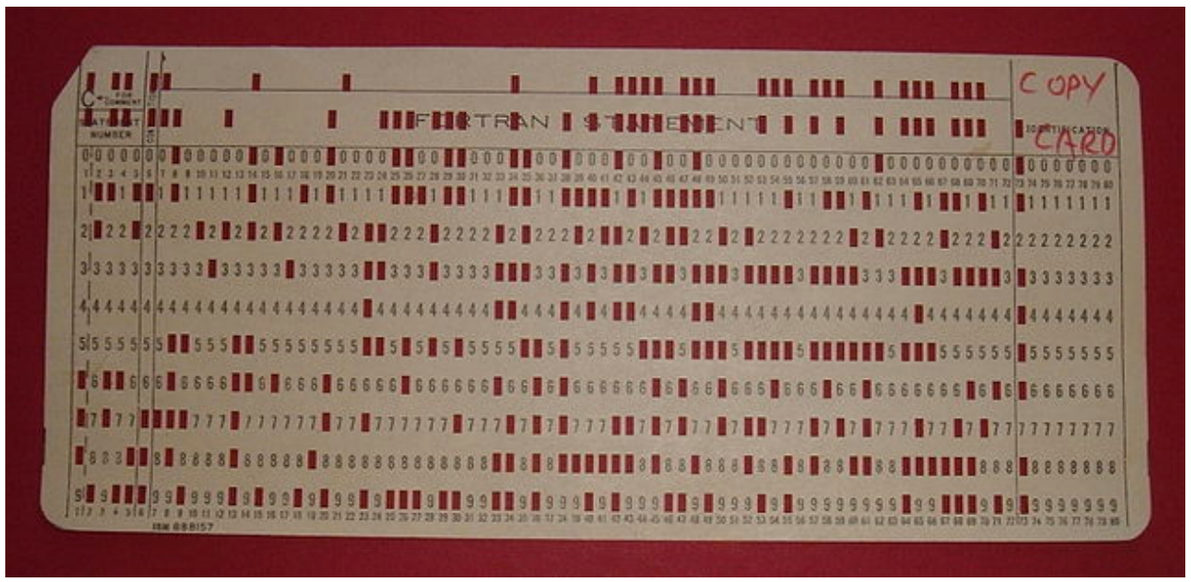
5. Hardware obsolescence



5. Hardware obsolescence



5. Hardware obsolescence



Staying stuck in a technology costs a lot in the mid-term

Hard disk evolution

IBM 305 RAMAC (1956)

eq. 27.882 USD in 2017

5 MB

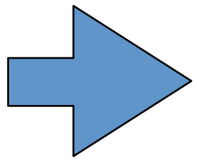
~2000 h MTBF

910 kg

Hard disk evolution

IBM 305 RAMAC (1956)

eq. 27.882 USD in 2017
5 MB
~2000 h MTBF
910 kg

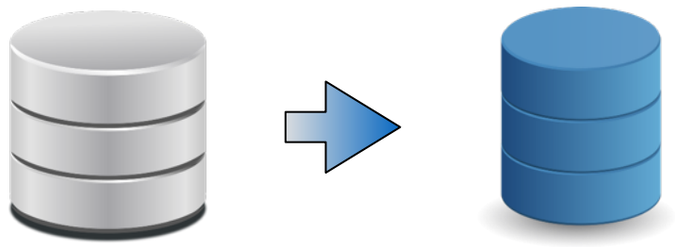


WD Gold (2017)

400 USD
12 To
2.5 M h MTBF
660 g

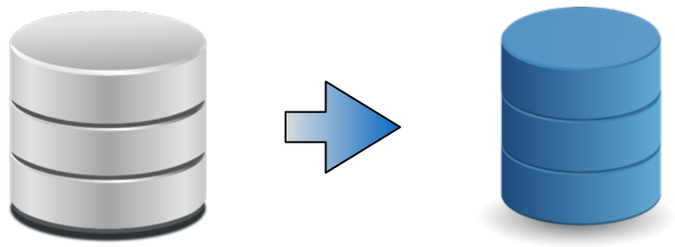
Need to migrate storage media to the most cost-effective technology

Regular migration to recent hardware technology with the best trade-off capacity/reliability/cost



Need to migrate storage media to the most cost-effective technology

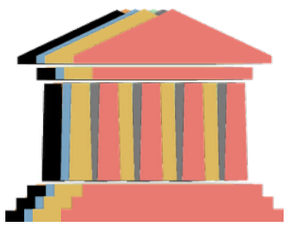
Regular migration to recent hardware technology with the best trade-off capacity/reliability/cost

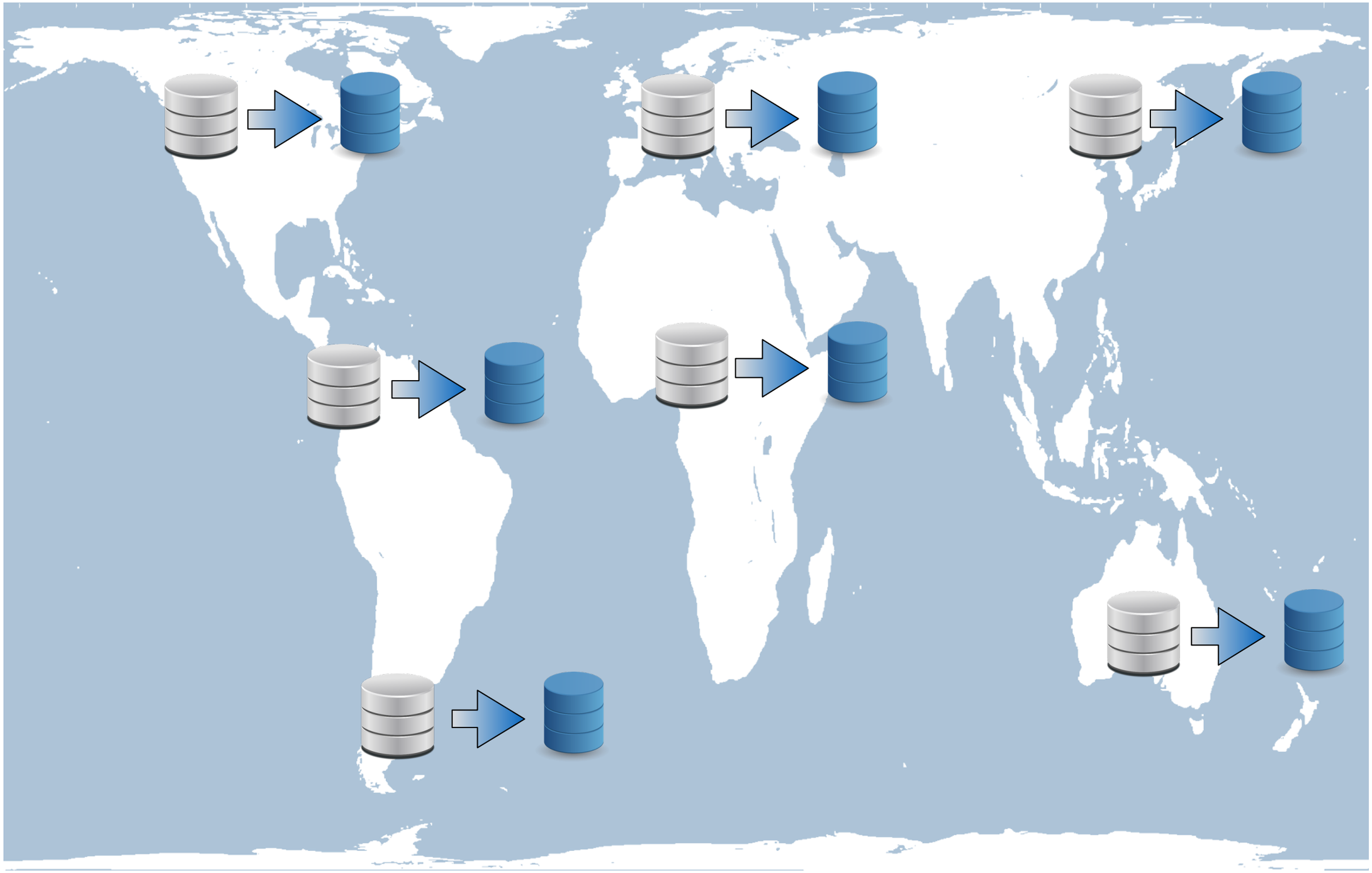


every 3 to 5 yrs

5. Hardware obsolescence

→ Media migration





6. Hardware failures

At the moment, hard disk drives is still the best choice

- low purchasing cost
- low maintenance cost
- reliable (MTBF in Mh)
- regular capacity evolution

WD Gold
Enterprise-class

400 USD
12 TB
2.5 M h MTBF
660 g

Complete hard disk failures are quite rare…

Hard Drive Annualized Failure Rates for Q4 2016

Reporting period 10/1/2016 - 12/31/2016 inclusive

MFG	Model	Drive Size	Drive Count	Avg. Age (months)	Drive Days	Drive Failures	Failure Rate
HGST	HDS723030ALA640	3 TB	978	61.21	90,415	9	3.63%
HGST	HDS5C3030ALA630	3 TB	4,476	55.87	412,752	13	1.15%
HGST	HDS5C4040ALE630	4 TB	2,625	45.35	241,665	4	0.60%
Toshiba	DT01ACA300	3 TB	46	44.12	4,232	-	0.00%
Seagate	ST4000DX000	4 TB	184	38.54	17,354	7	14.72%
WDC	WD30EFRX	3 TB	1,105	30.39	100,259	9	3.28%
HGST	HMS5C4040ALE640	4 TB	7,014	29.48	648,393	9	0.51%
WDC	WD60EFRX	6 TB	446	24.14	41,304	5	4.42%
HGST	HUH728080ALE600	8 TB	45	22.99	4,140	-	0.00%
Toshiba	MD04ABA500V	5 TB	45	22.15	4,140	-	0.00%
Seagate	ST4000DM000	4 TB	34,738	21.73	3,196,552	234	2.67%
Seagate	ST6000DX000	6 TB	1,889	21.48	173,720	8	1.68%
Toshiba	MD04ABA400V	4 TB	146	20.61	13,432	-	0.00%
WDC	WD40EFRX	4 TB	75	17.16	4,232	-	0.00%
HGST	HMS5C4040BLE640	4 TB	9,407	15.51	809,119	14	0.63%
Seagate	ST8000DM002	8 TB	8,660	4.72	663,697	30	1.65%
Seagate	ST8000NM0055	8 TB	60	1.44	1,560	-	0.00%
		Totals	**71,939**		**6,426,966**	**342**	**1.94%**

https://www.backblaze.com/blog/hard-drive-benchmark-stats-2016/

BACKBLAZE

Complete hard disk failures are quite rare…

Hard Drive Annualized Failure Rates for Q4 2016

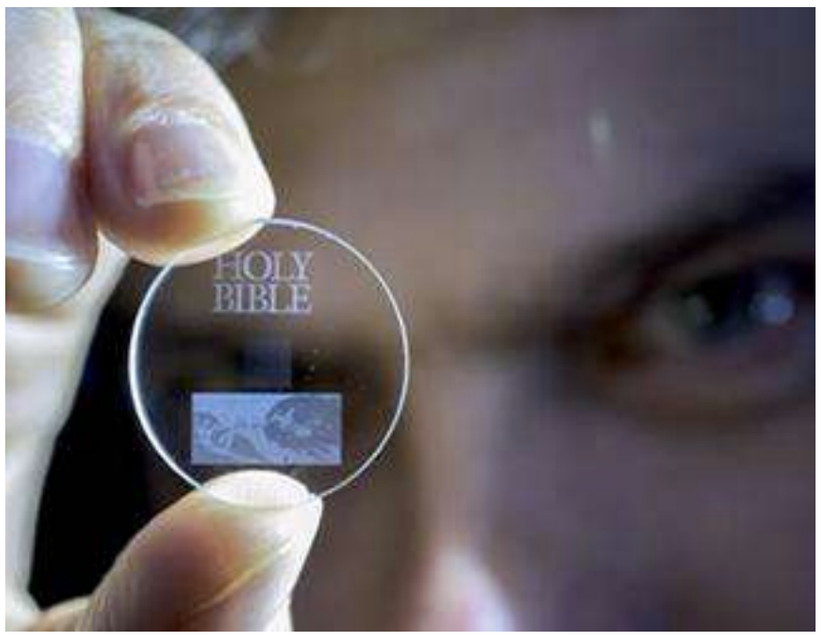
Reporting period 10/1/2016 - 12/31/2016 inclusive

MFG	Model	Drive Size	Drive Count	Avg. Age (months)	Drive Days	Drive Failures	Failure Rate
HGST	HDS723030ALA640	3 TB	978	61.21	90,415	9	3.63%
HGST	HDS5C3030ALA630	3 TB	4,476	55.87	412,752	13	1.15%
HGST	HDS5C4040ALE630	4 TB	2,625	45.35	241,665	4	0.60%
Toshiba	DT01ACA300	3 TB	46	44.12	4,232	-	0.00%
Seagate	ST4000DX000	4 TB	184	38.54	17,354	7	14.72%
WDC	WD30EFRX	3 TB	1,105	30.39	100,259	9	3.28%
HGST	HMS5C4040ALE640	4 TB	7,014	29.48	648,393	9	0.51%
WDC	WD60EFRX	6 TB	446	24.14	41,304	5	4.42%
HGST	HUH728080ALE600	8 TB	45	22.99	4,140	-	0.00%
Toshiba	MD04ABA500V	5 TB	45	22.15	4,140	-	0.00%
Seagate	ST4000DM000	4 TB	34,738	21.73	3,196,552	234	2.67%
Seagate	ST6000DX000	6 TB	1,889	21.48	173,720	8	1.68%
Toshiba	MD04ABA400V	4 TB	146	20.61	13,432	-	0.00%
WDC	WD40EFRX	4 TB	75	17.16	4,232	-	0.00%
HGST	HMS5C4040BLE640	4 TB	9,407	15.51	809,119	14	0.63%
Seagate	ST8000DM002	8 TB	8,660	4.72	663,697	30	1.65%
Seagate	ST8000NM0055	8 TB	60	1.44	1,560	-	0.00%
		Totals	**71,939**		**6,426,966**	**342**	**1.94%**

https://www.backblaze.com/blog/hard-drive-benchmark-stats-2016/

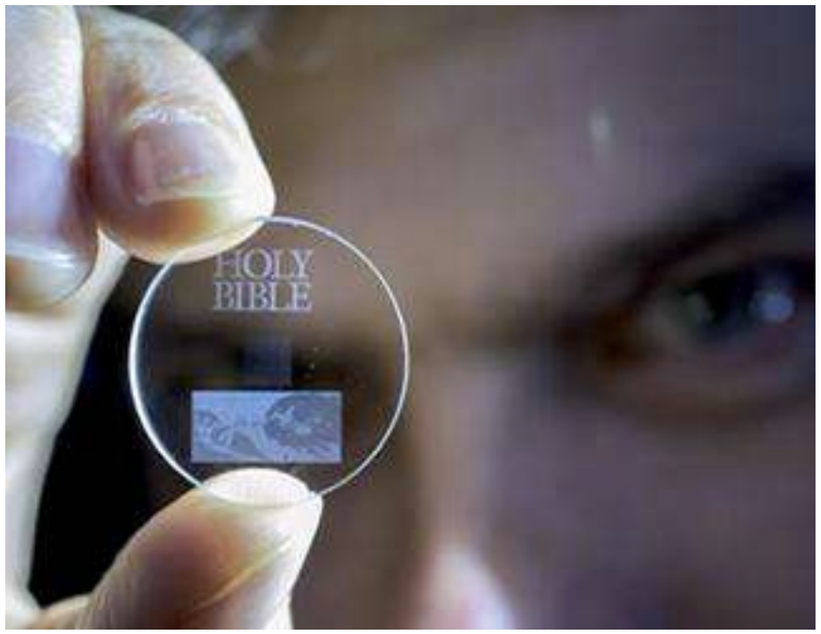
BACKBLAZE

Super media

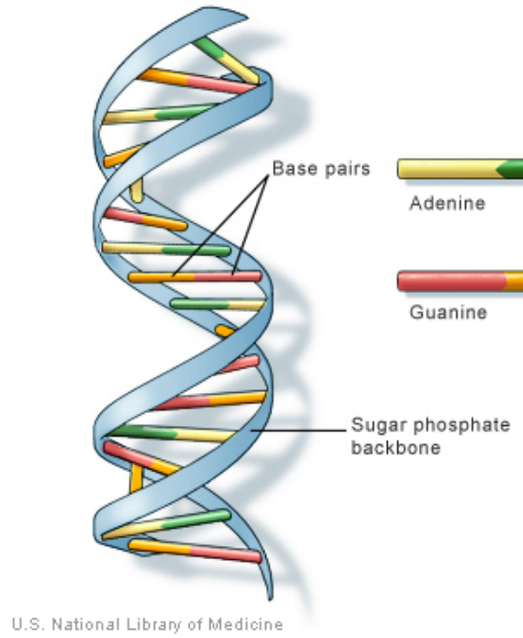


“Eternal 5D” Storage
(>13 Mds yrs)

Super media

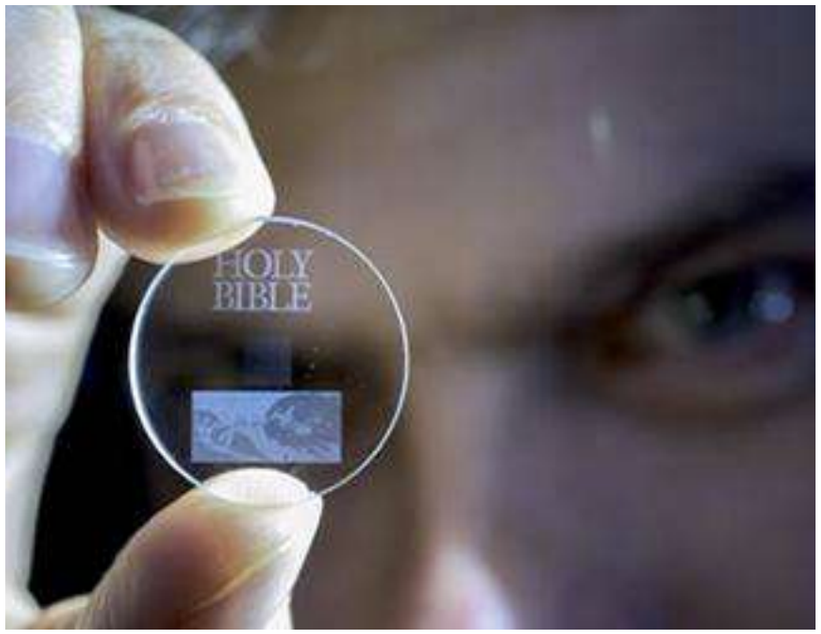


“Eternal 5D” Storage
(>13 Mds yrs)

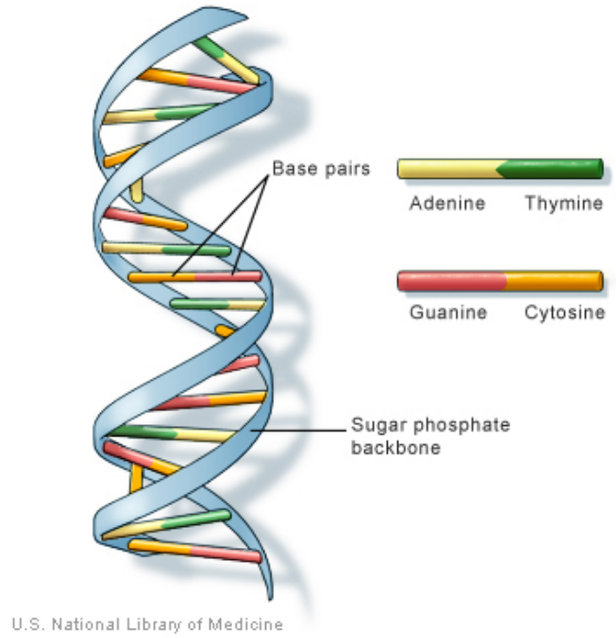


DNA Storage
7.5 M$ / GB

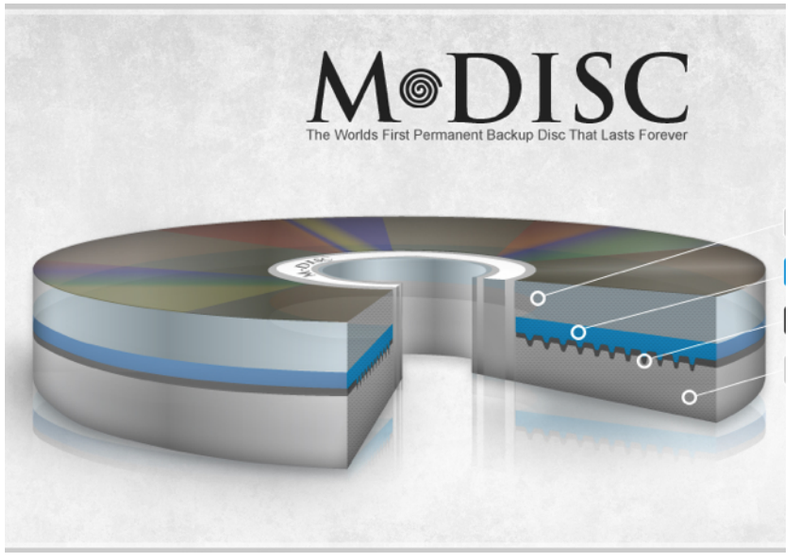
Super media



“Eternal 5D” Storage
(>13 Mds yrs)

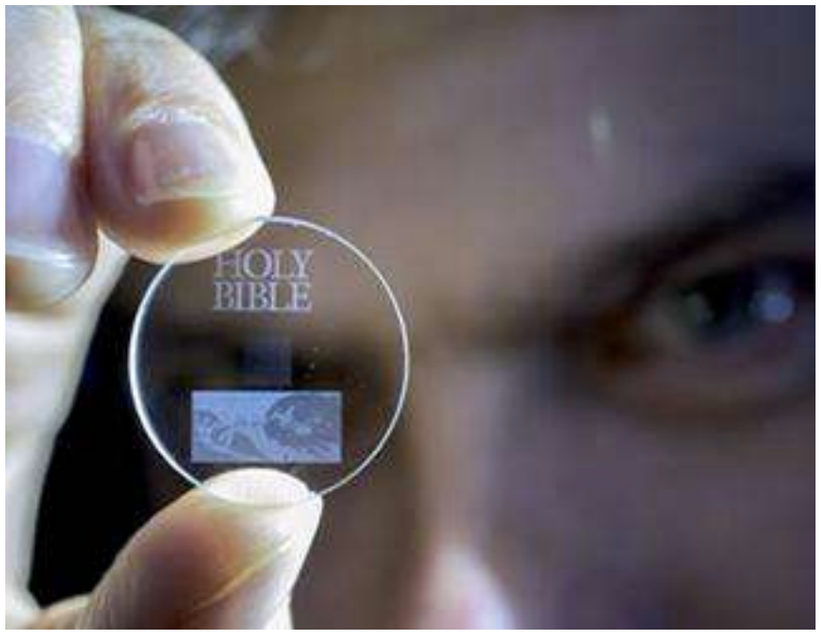


DNA Storage
7.5 M$ / GB

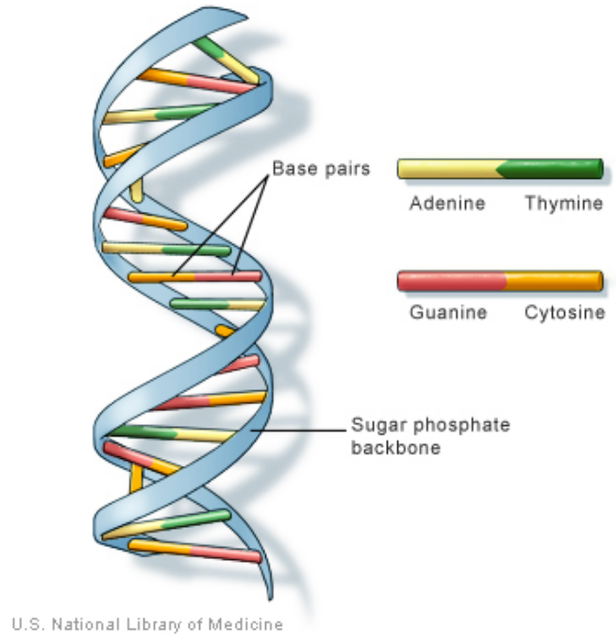


M disc (1000 yrs)

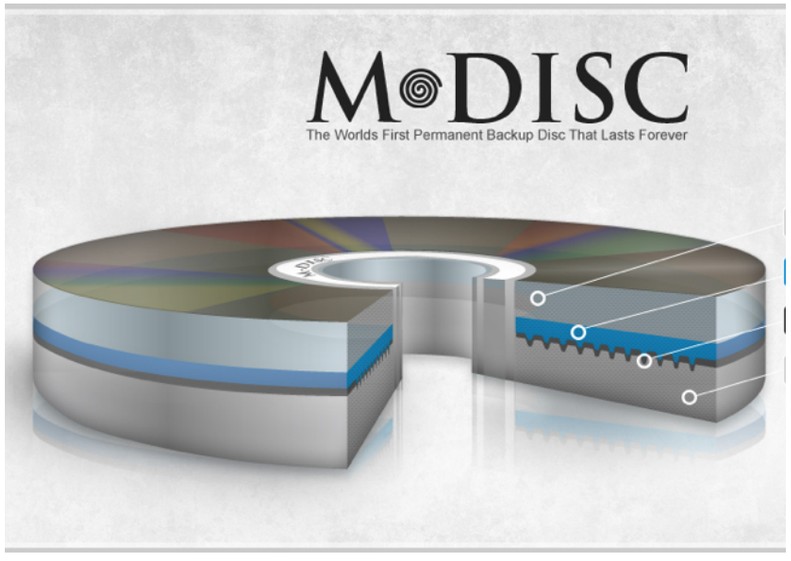
Super media



“Eternal 5D” Storage
(>13 Mds yrs)



DNA Storage
7.5 M$ / GB



M disc (1000 yrs)

Problems (for now) :
- economic viability still to be confirmed
- difficult to perform integrity monitoring

The “bit-rot” or data degradation

Magnetic storage

Solid-state storage

With the increase of storage capacity, the risk of accidental data degradation increase.

The “bit-rot” or data degradation

Magnetic storage

Solid-state storage

With the increase of storage capacity, the risk of accidental data degradation increase.

Solution: check and rewrite data regularly

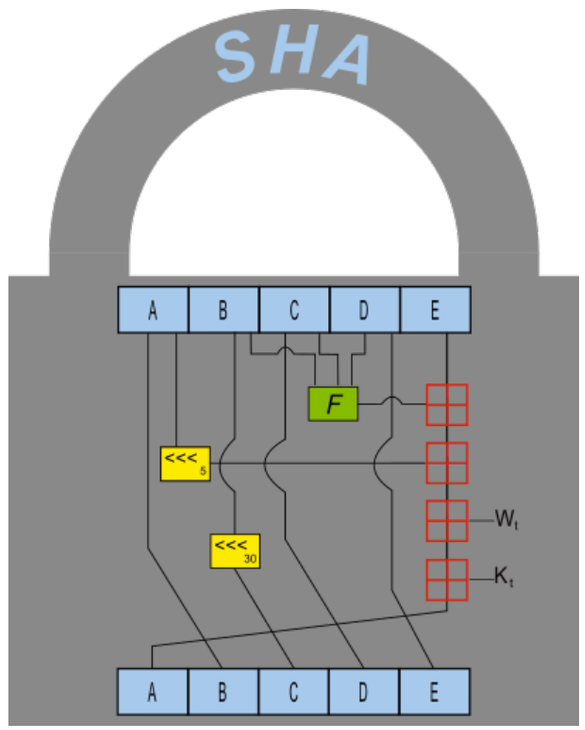
How to ensure the copies integrity against data degradation?

Université libre de Bruxelles

How to ensure the copies integrity against data degradation?

checksum (hash)

Université libre de Bruxelles → 4e8b8da4a7315fd8b5e5dcc5ac313f6657d730c4

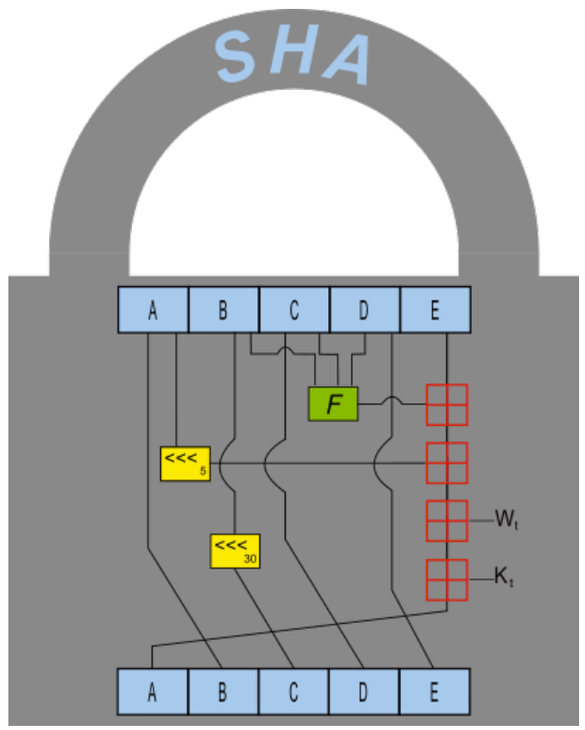


How to ensure the copies integrity against data degradation?

checksum (hash)

Université **l**ibre de Bruxelles → 4e8b8da4a7315fd8b5e5dcc5ac313f6657d730c4

Université **L**ibre de Bruxelles

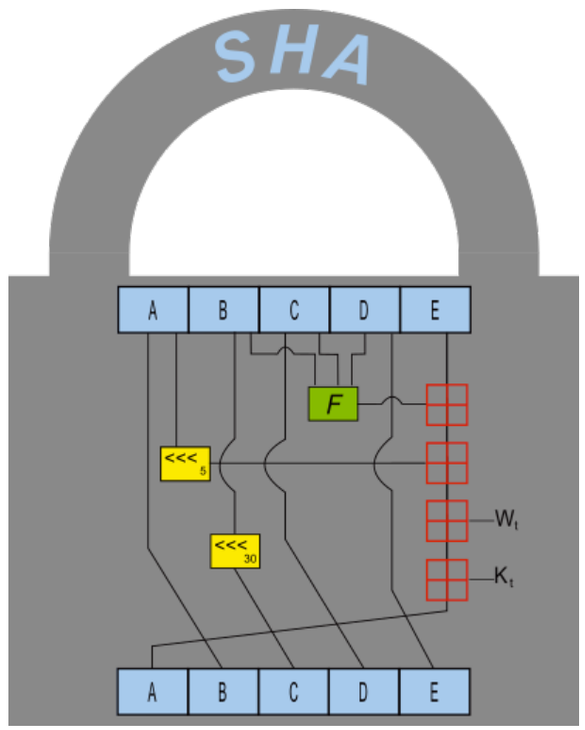


How to ensure the copies integrity against data degradation?

checksum (hash)

Université **l**ibre de Bruxelles → 4e8b8da4a7315fd8b5e5dcc5ac313f6657d730c4

Université **L**ibre de Bruxelles → bddef72a8beb5aab6b199fac36b1274ba77ecb06



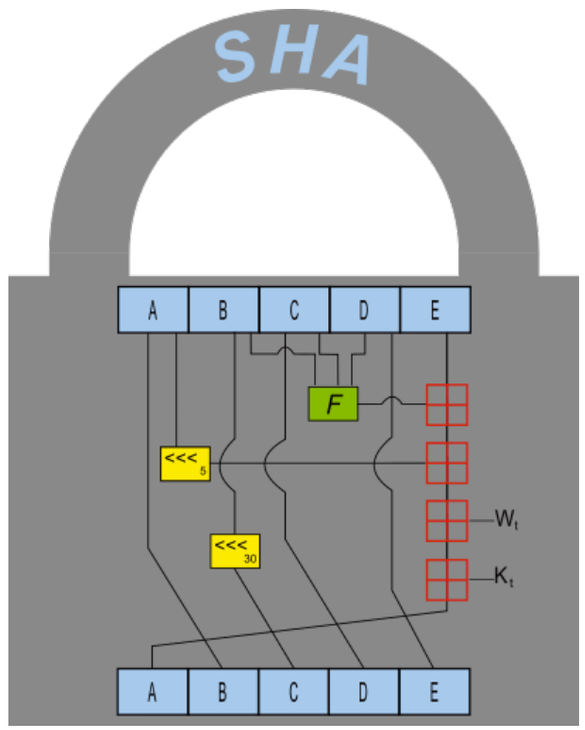
How to ensure the copies integrity against data degradation?

checksum (hash)

Université **l**ibre de Bruxelles → 4e8b8da4a7315fd8b5e5dcc5ac313f6657d730c4

Université **L**ibre de Bruxelles → bddef72a8beb5aab6b199fac36b1274ba77ecb06

Effective integrity check to detect **accidental errors**.



We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value

Original Checksum

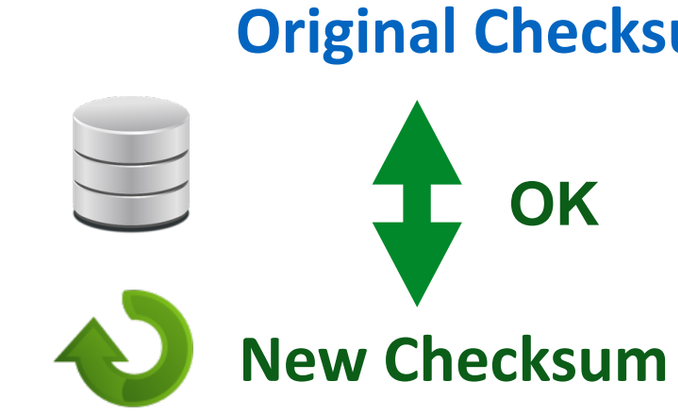
We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value

Original Checksum

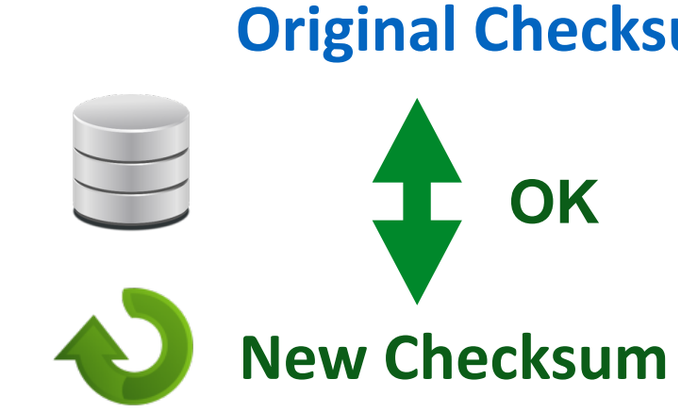
New Checksum

regularly

We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value

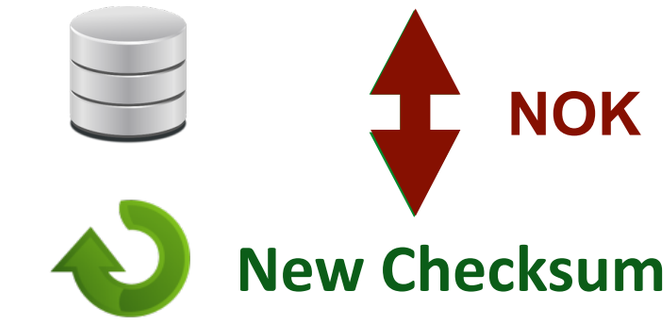


We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value



Problem :
The checksums are prone to the same threats than the data we want to preserve !

We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value

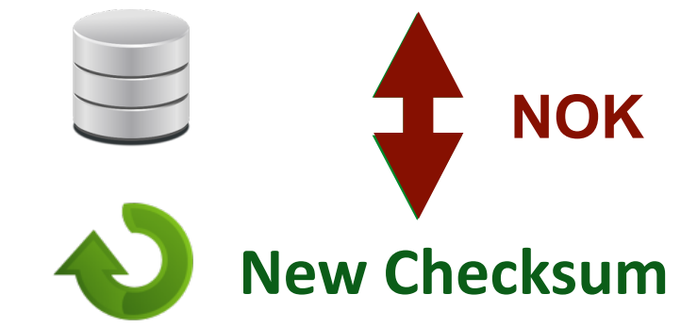


Problem :
The checksums are prone to the same threats than the data we want to preserve !

If the new hash does not match with the original hash:

- either the object is corrupted
- or the checksum is corrupted
- or both are corrupted

We need to compute a checksum at submission and regularly reevaluate the checksum and compare it to the original value



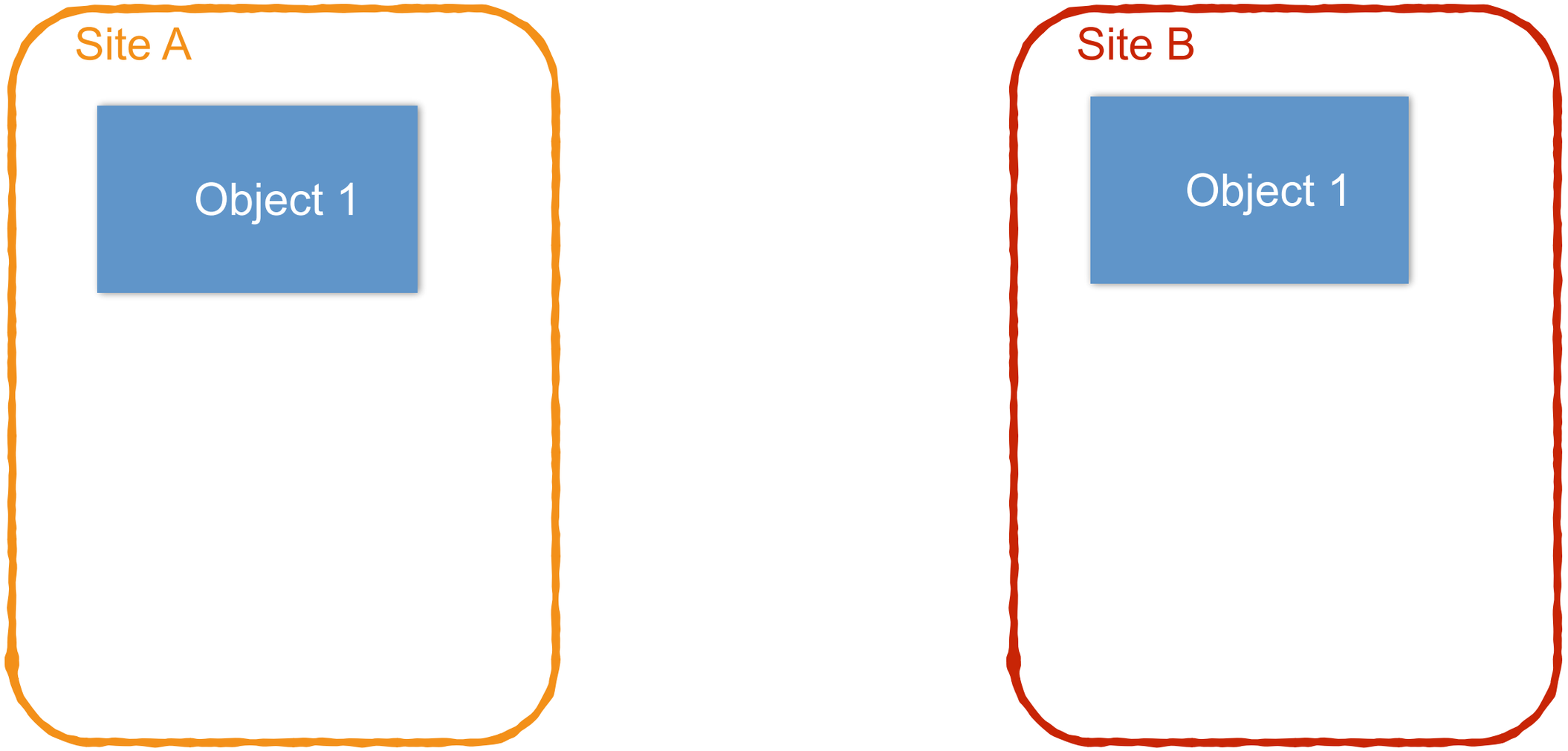
Problem :
The checksums are prone to the same threats than the data we want to preserve !

If the new hash does not match with the original hash:

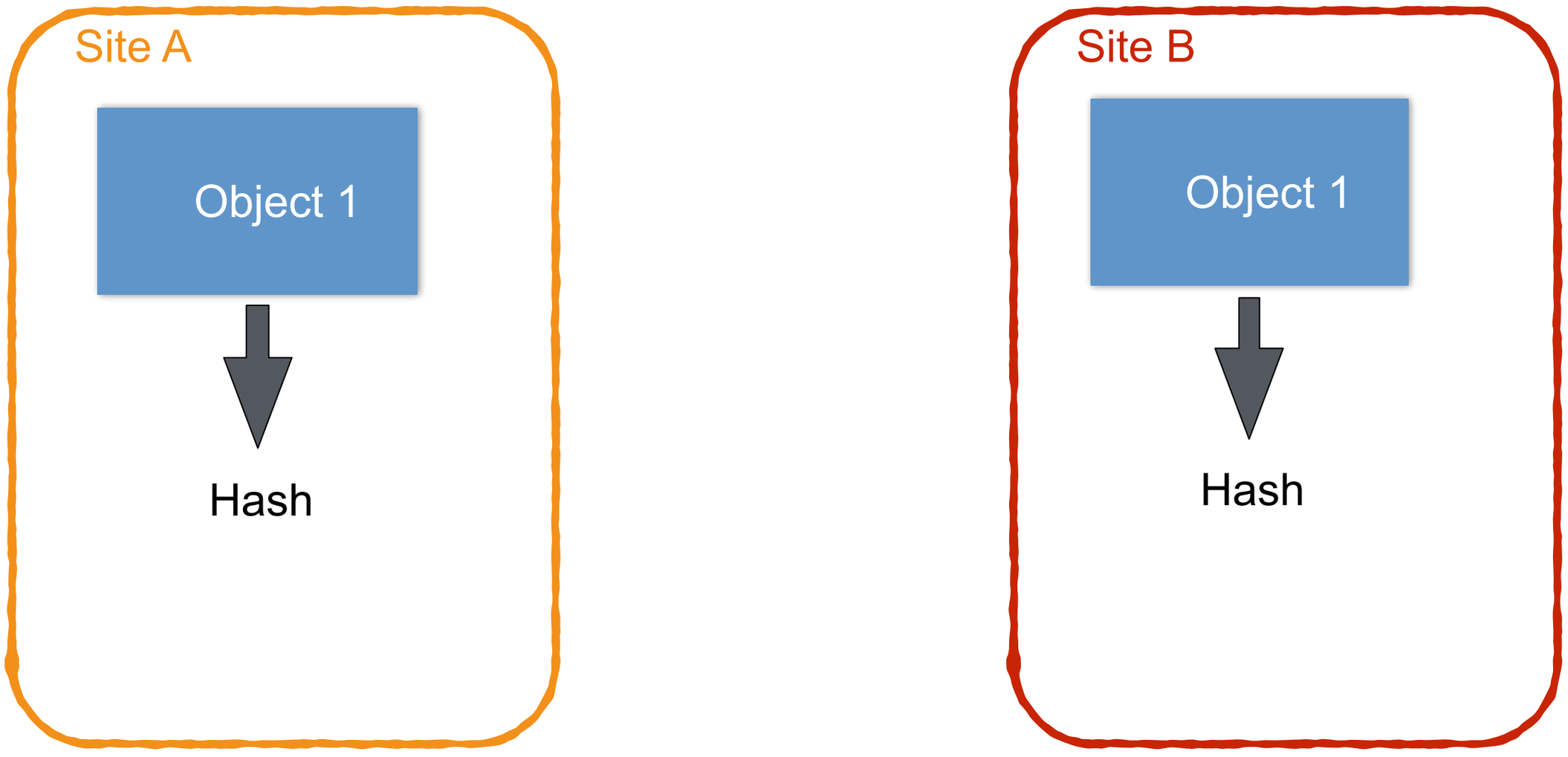
- either the object is corrupted
- or the checksum is corrupted
- or both are corrupted

It would be safer to compare the new checksum with the checksums of the other copies

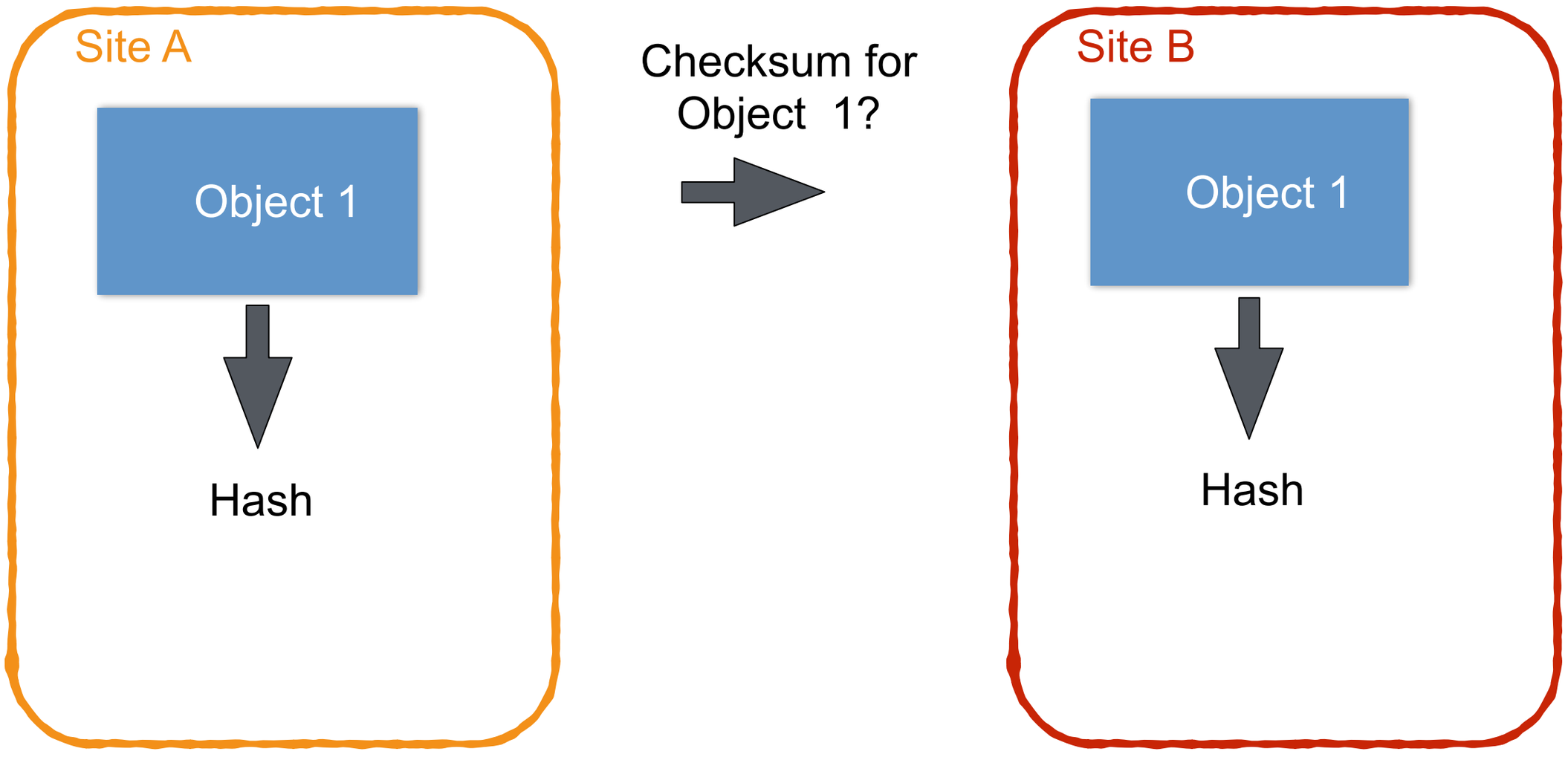
Comparison with other copies in the network



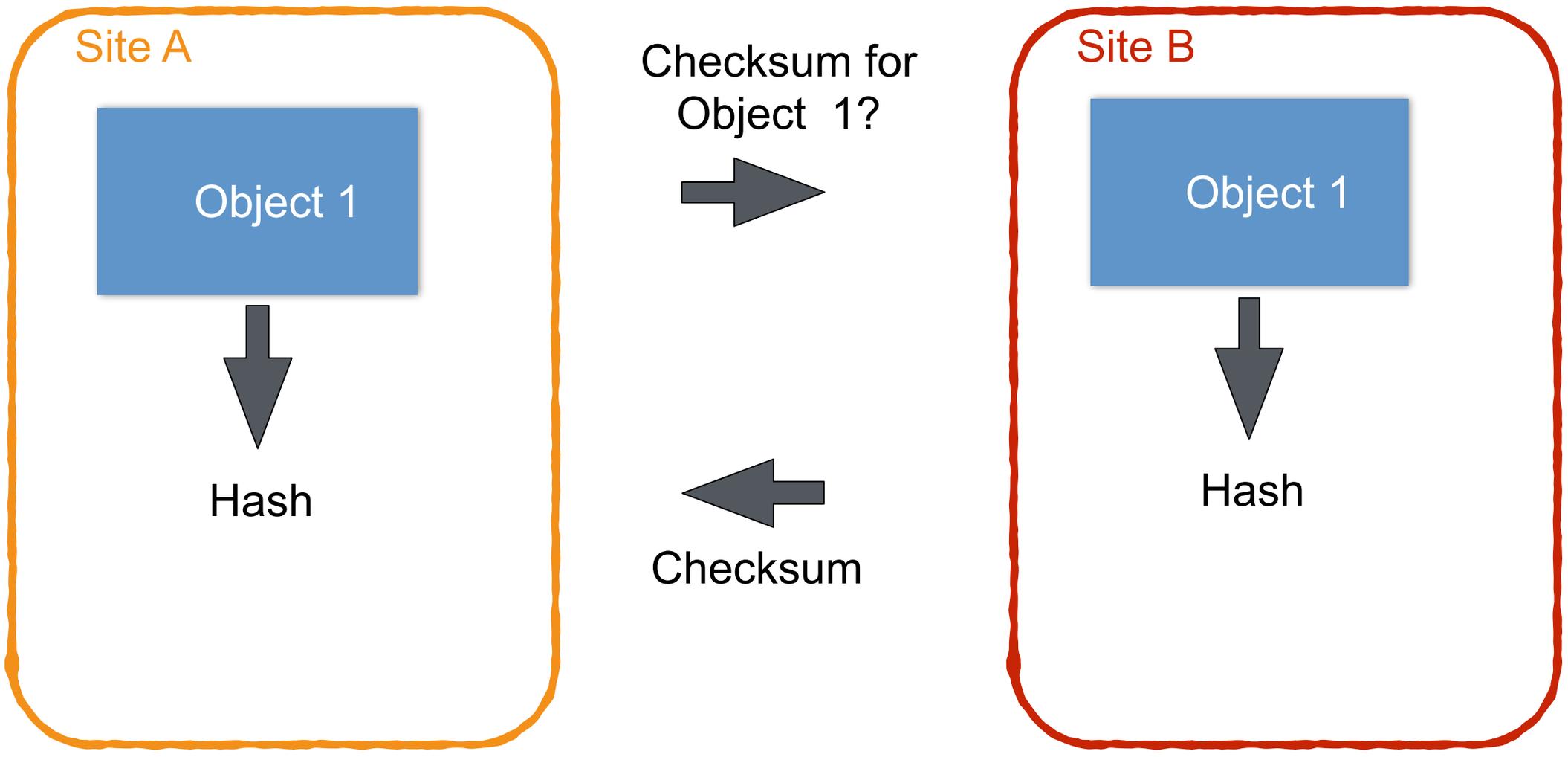
Comparison with other copies in the network



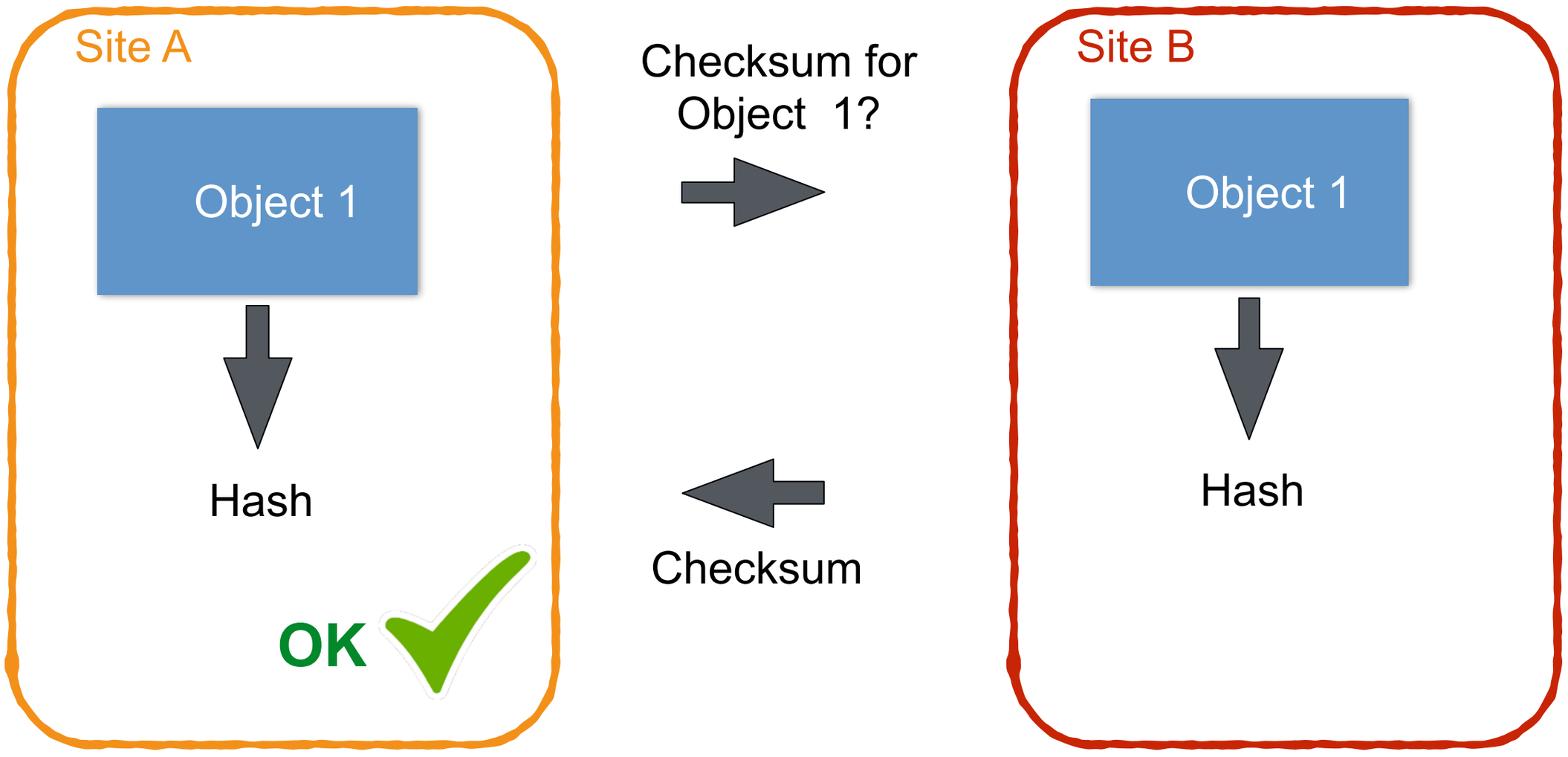
Comparison with other copies in the network



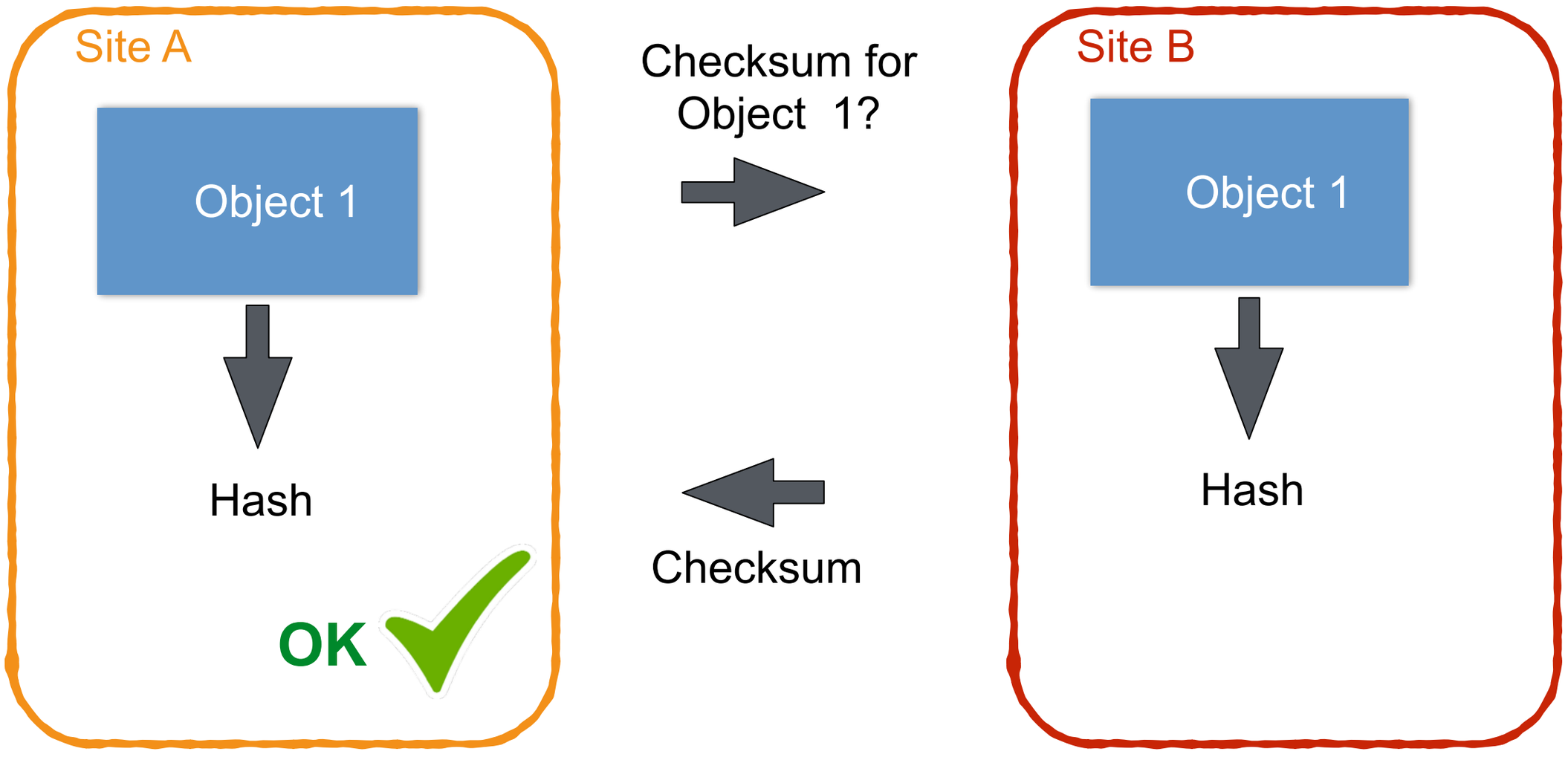
Comparison with other copies in the network



Comparison with other copies in the network



Comparison with other copies in the network



If we make the comparison with many copies, we considerably lower the probability of corruption.

6. Hardware failures

→ Regular data integrity monitoring

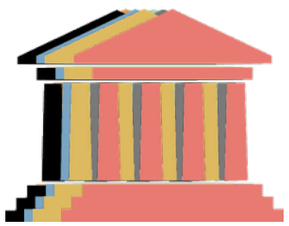
Regular check of the data integrity

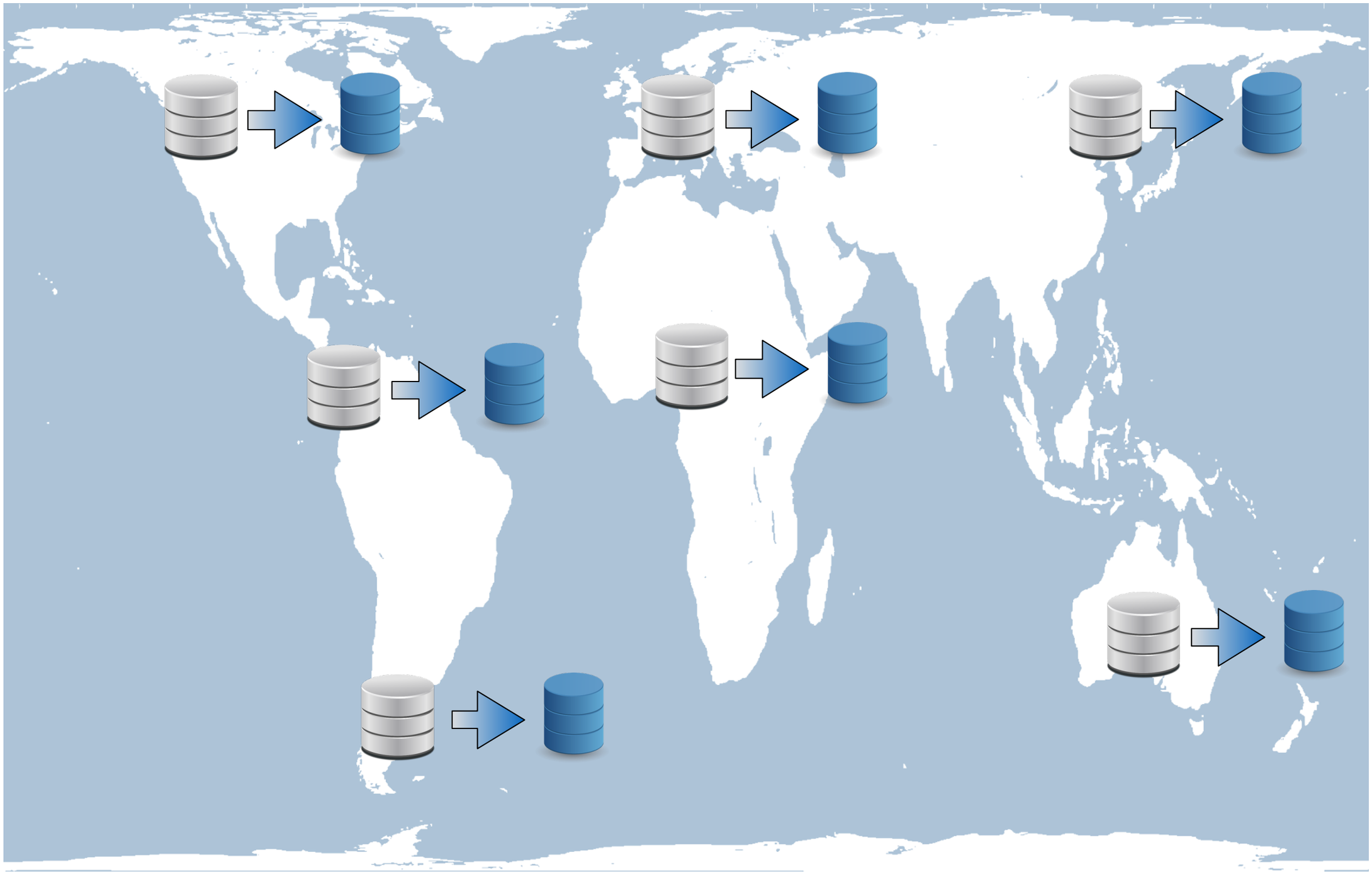
Replacing storage media regularly and at first signs of failure

Hardware heterogeneity to avoid failure correlation

6. Hardware failures

→ Regular data monitoring





7. External or internal attacks

http://www.doncio.navy.mil/chips/ArticleDetails.aspx?ID=9569

7. External or internal attacks

http://www.doncio.navy.mil/chips/ArticleDetails.aspx?ID=9569

7. External or internal attacks

http://www.doncio.navy.mil/chips/ArticleDetails.aspx?ID=9569

Security measures :

- Regular software updates
- Authentication
- Firewall

External attack : intentional corruption

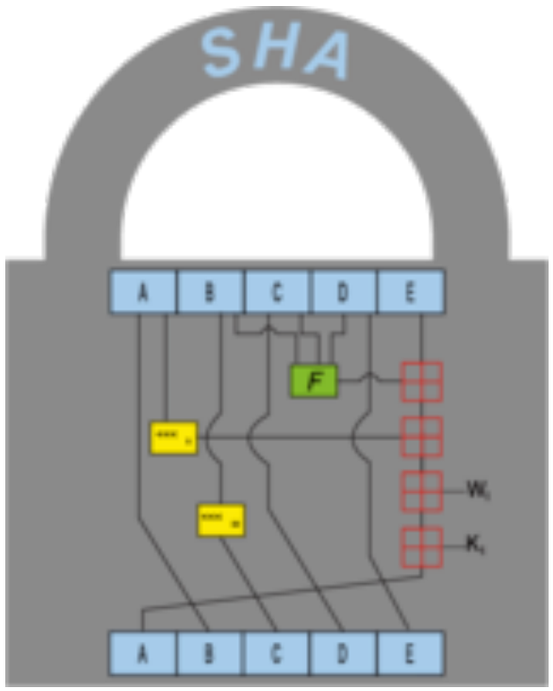
If an attacker wants to modify **on purpose** the content, they just need to adapt the checksums to the desired content.

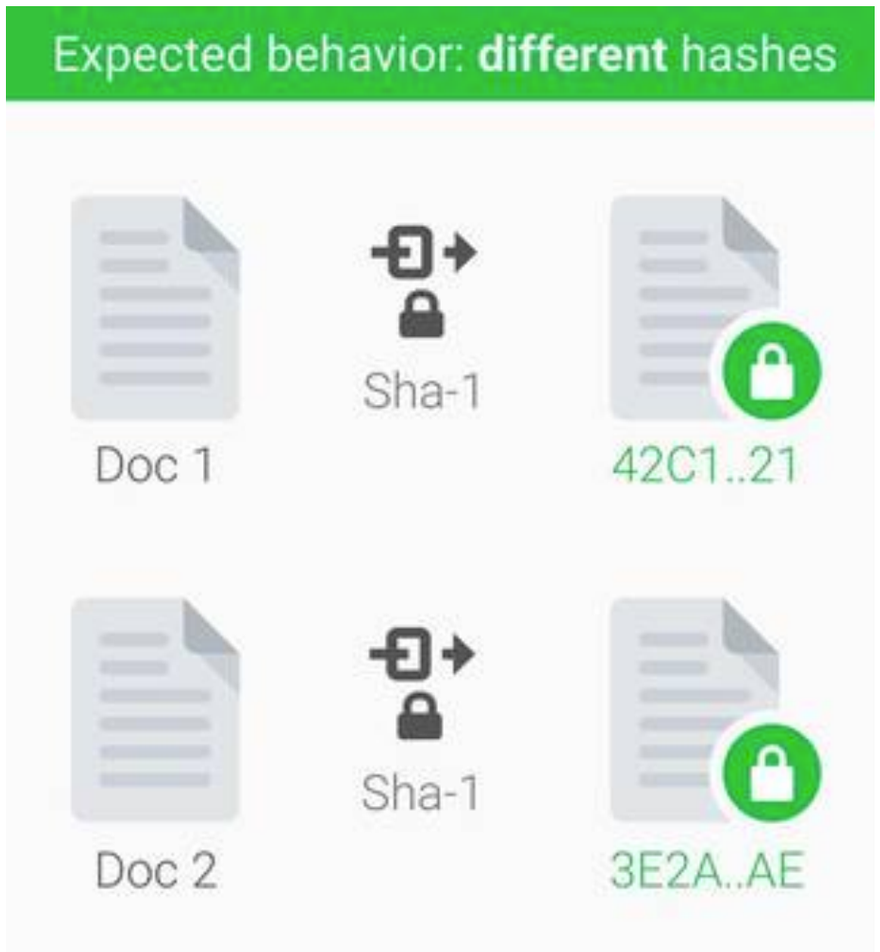
Marc Stevens, Elie Bursztein, Pierre Karpman, Ange Albertini, Yarik Markov, The first collision for full SHA-1, https://shattered.io/

External attack : intentional corruption

If an attacker wants to modify **on purpose** the content, they just need to adapt the checksums to the desired content.

More effective, checksum collisions are possible :



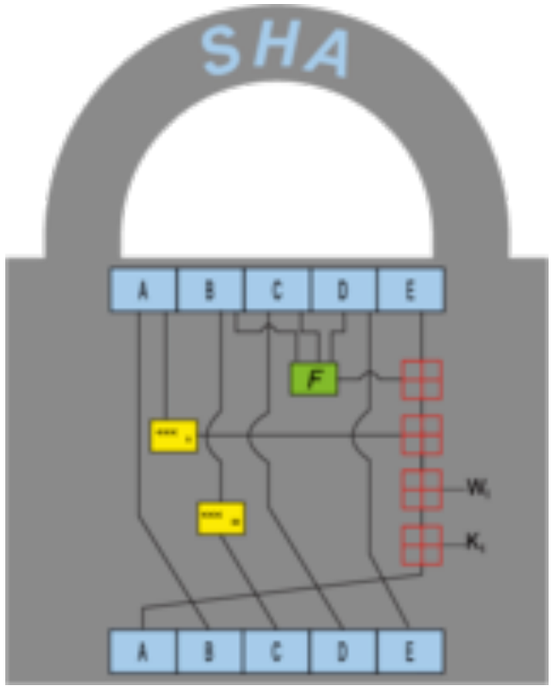


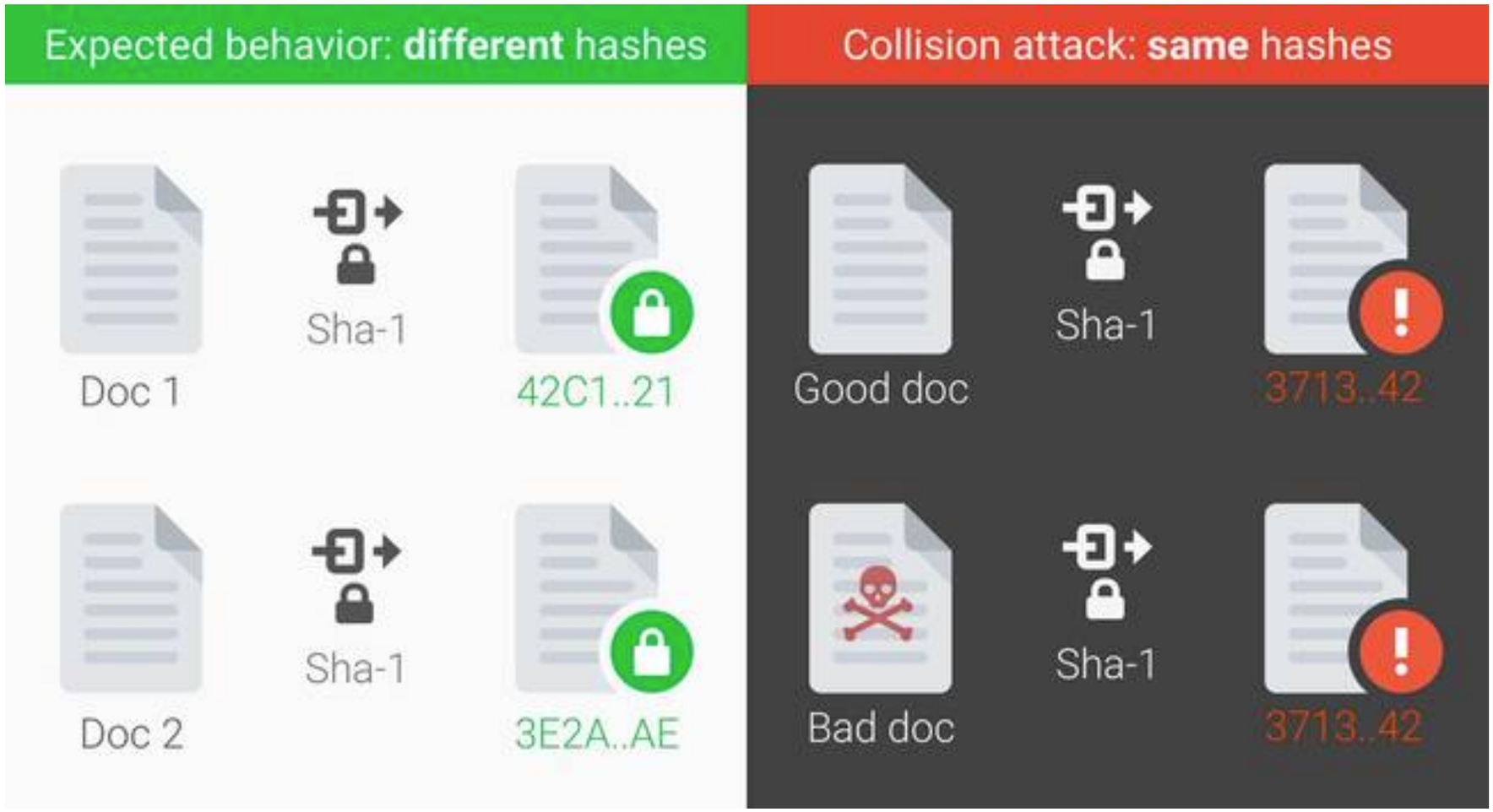
Marc Stevens, Elie Bursztein, Pierre Karpman, Ange Albertini, Yarik Markov, The first collision for full SHA-1, https://shattered.io/

External attack : intentional corruption

If an attacker wants to modify **on purpose** the content, they just need to adapt the checksums to the desired content.

More effective, checksum collisions are possible :

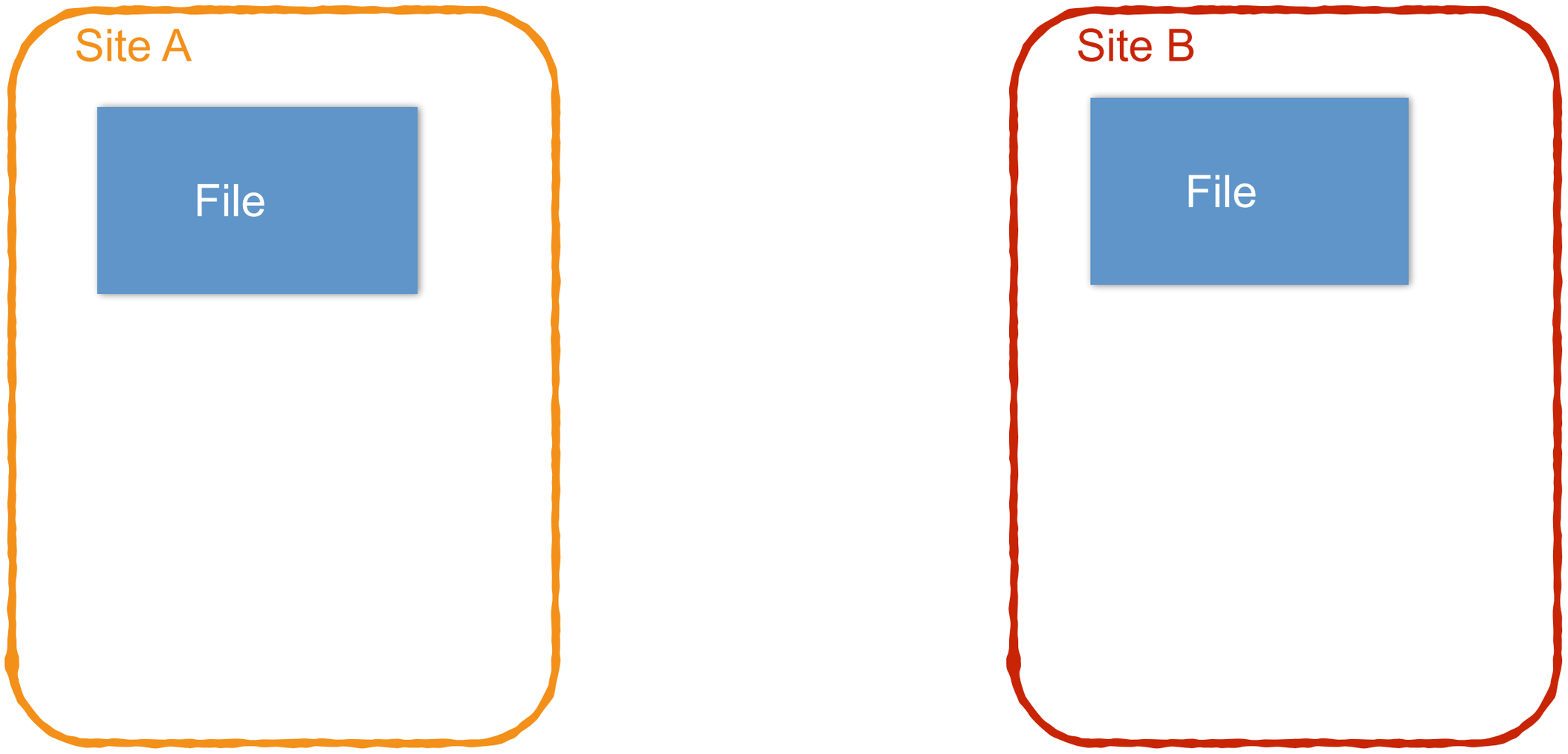




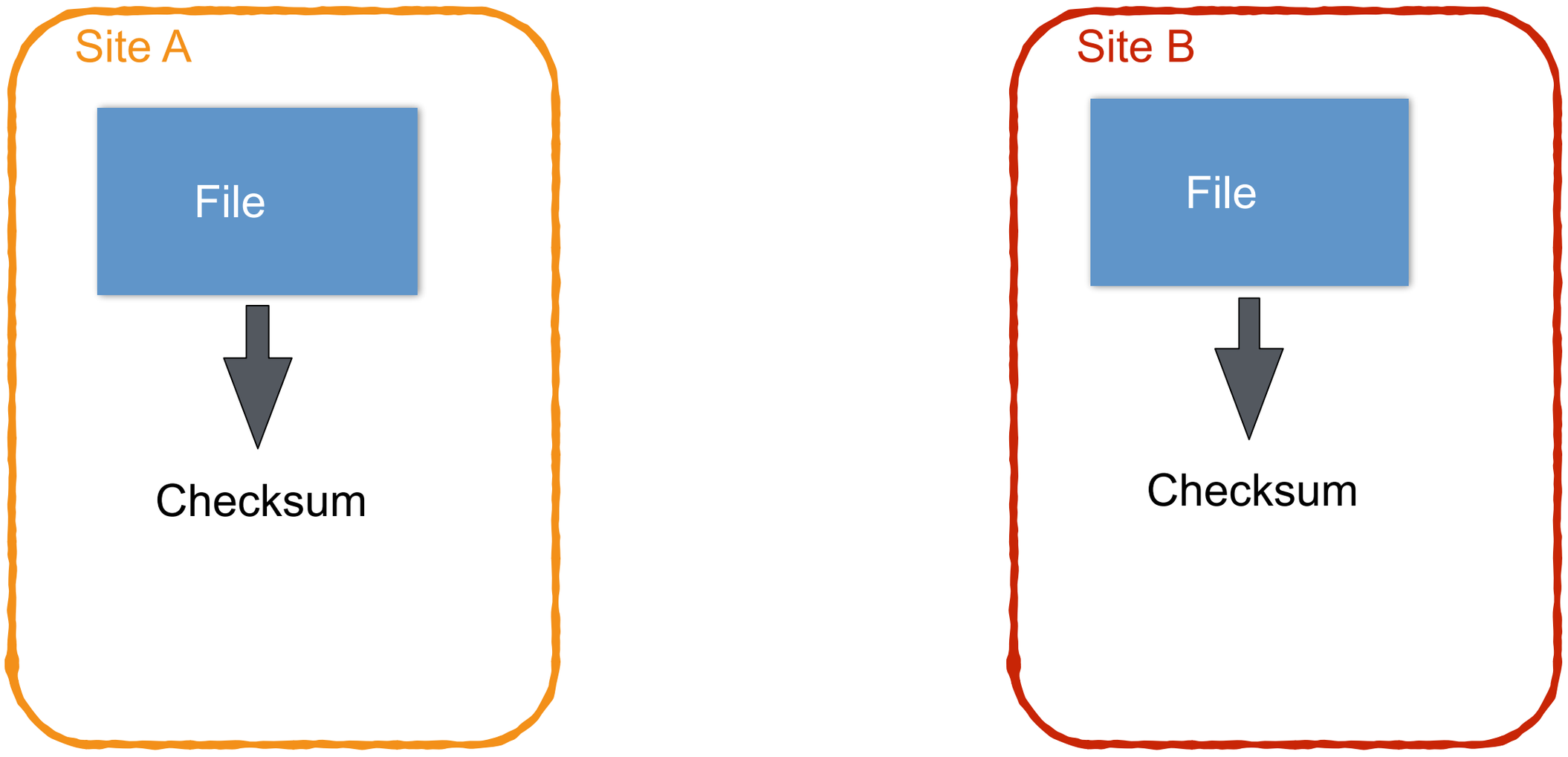
© Google

Marc Stevens, Elie Bursztein, Pierre Karpman, Ange Albertini, Yarik Markov, The first collision for full SHA-1, https://shattered.io/

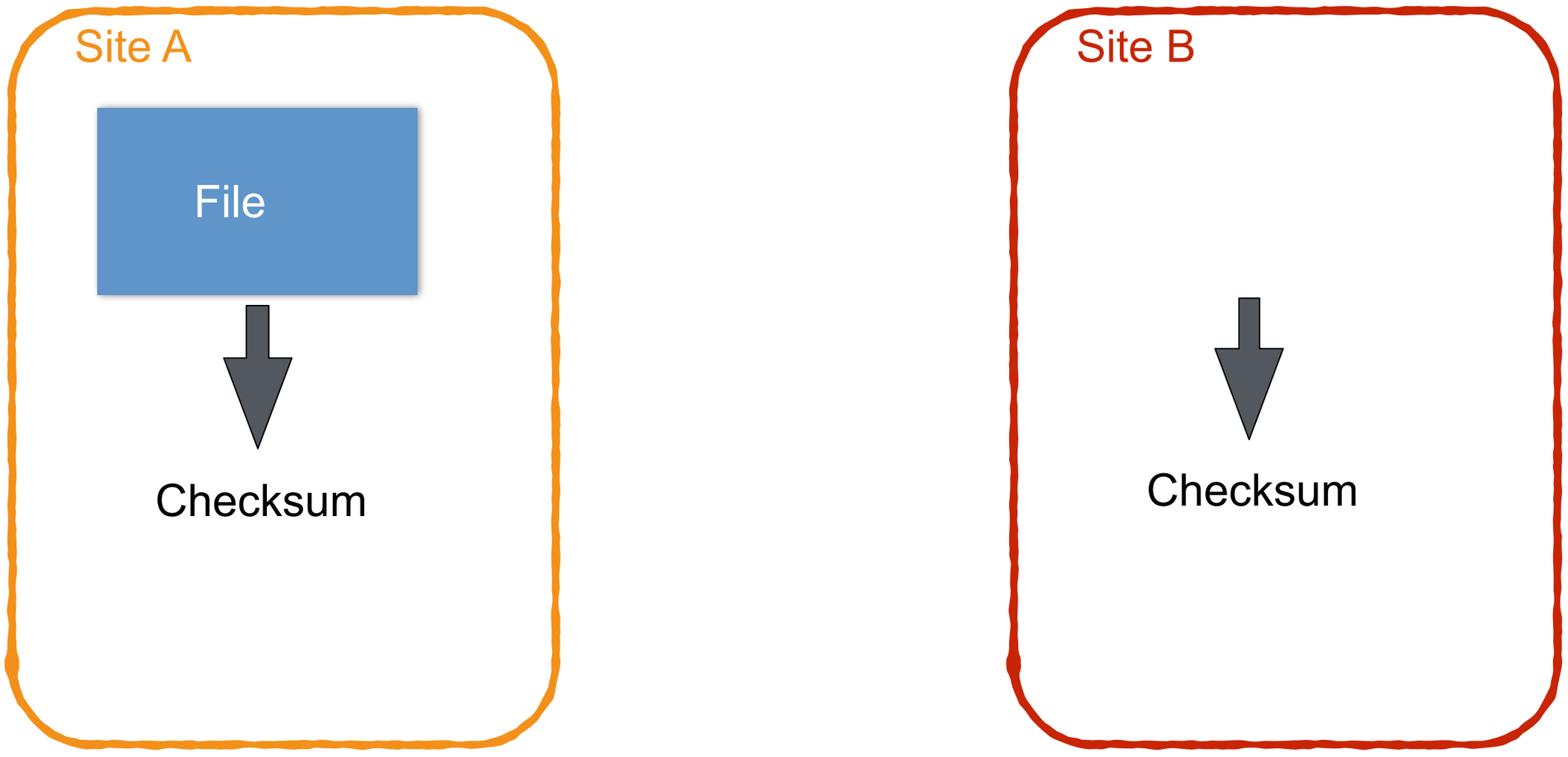
Internal attack : how can we ensure that we really have N copies in the network?



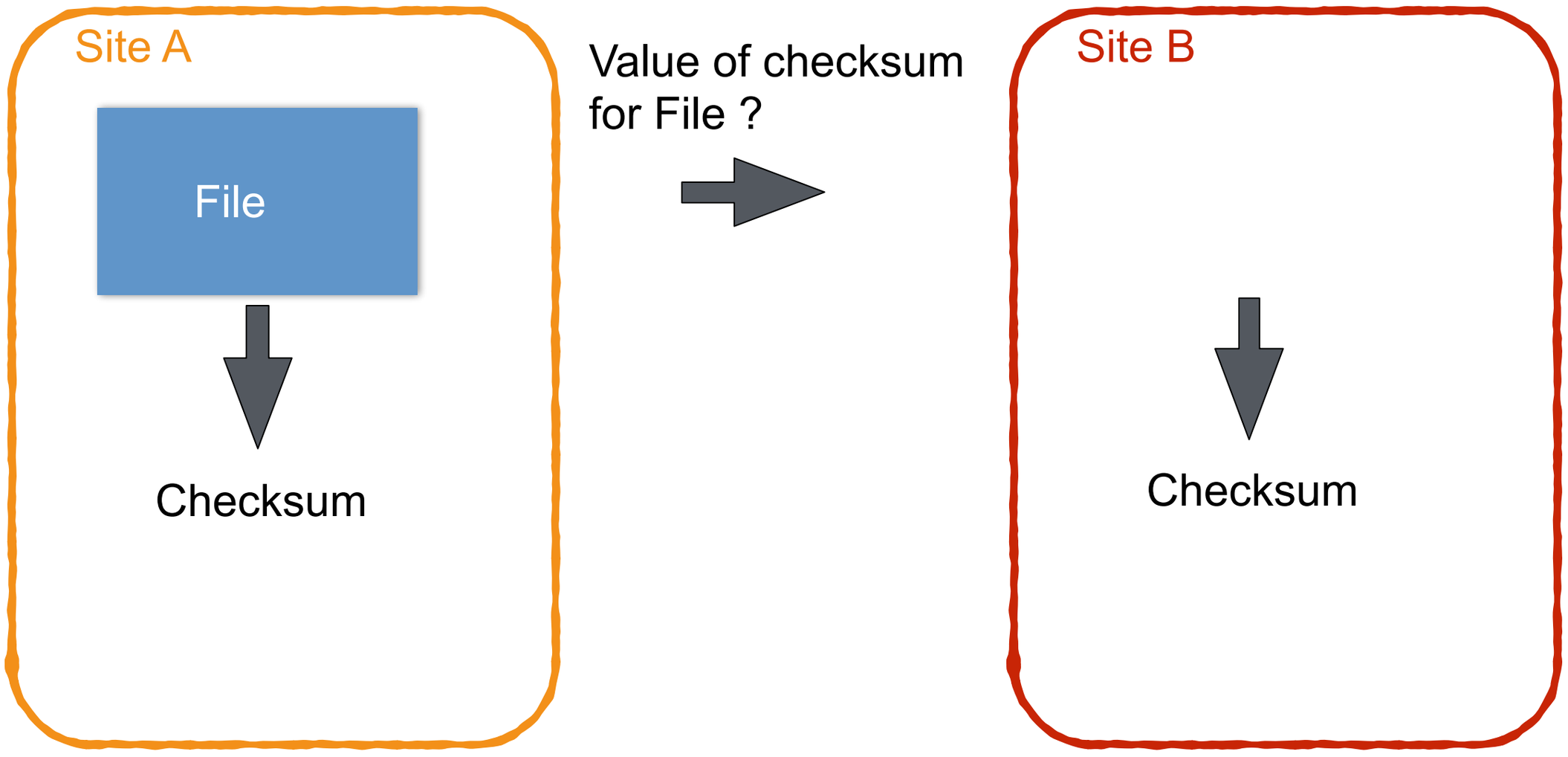
Internal attack : how can we ensure that we really have N copies in the network?



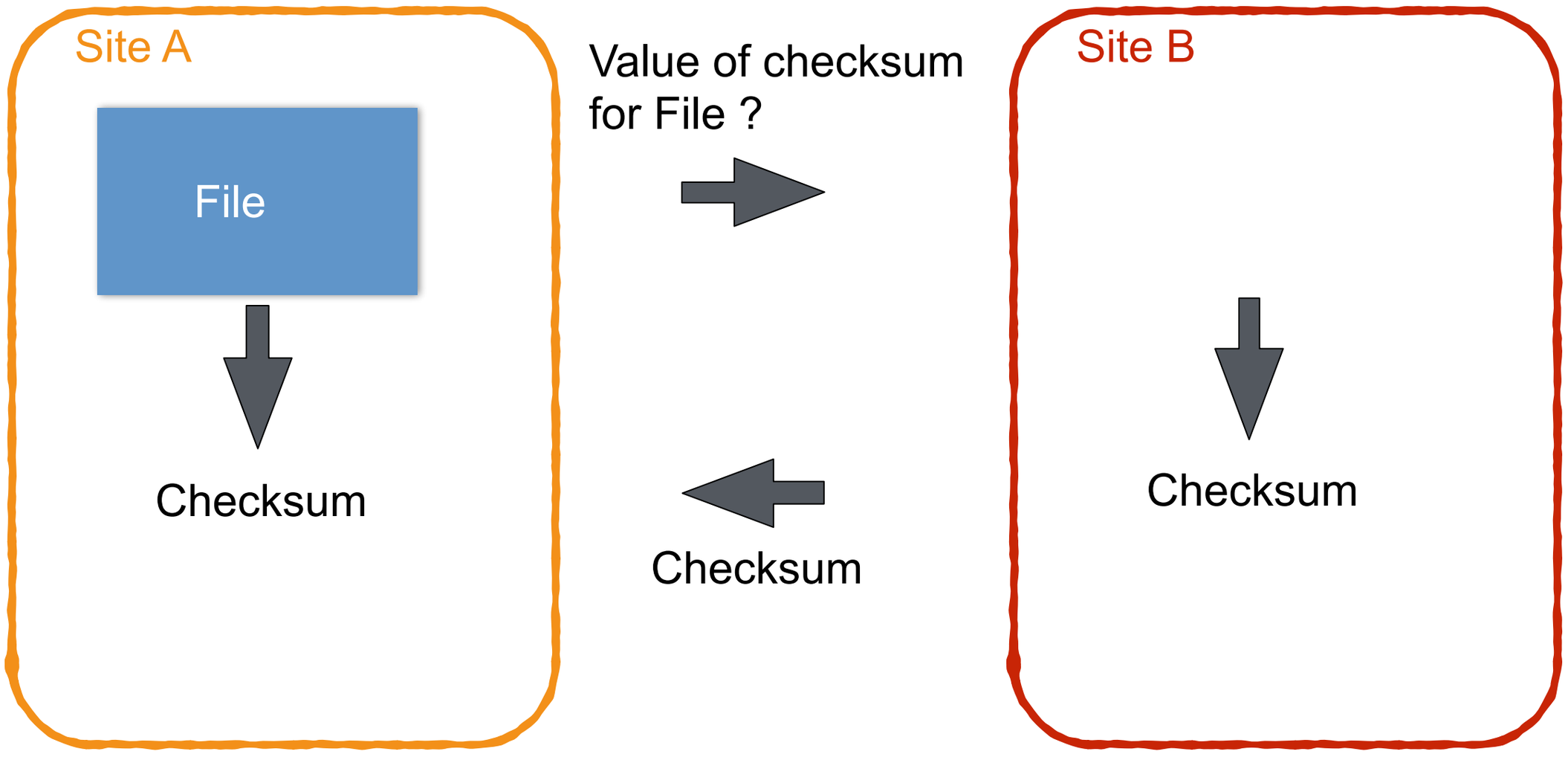
Internal attack : how can we ensure that we really have N copies in the network?



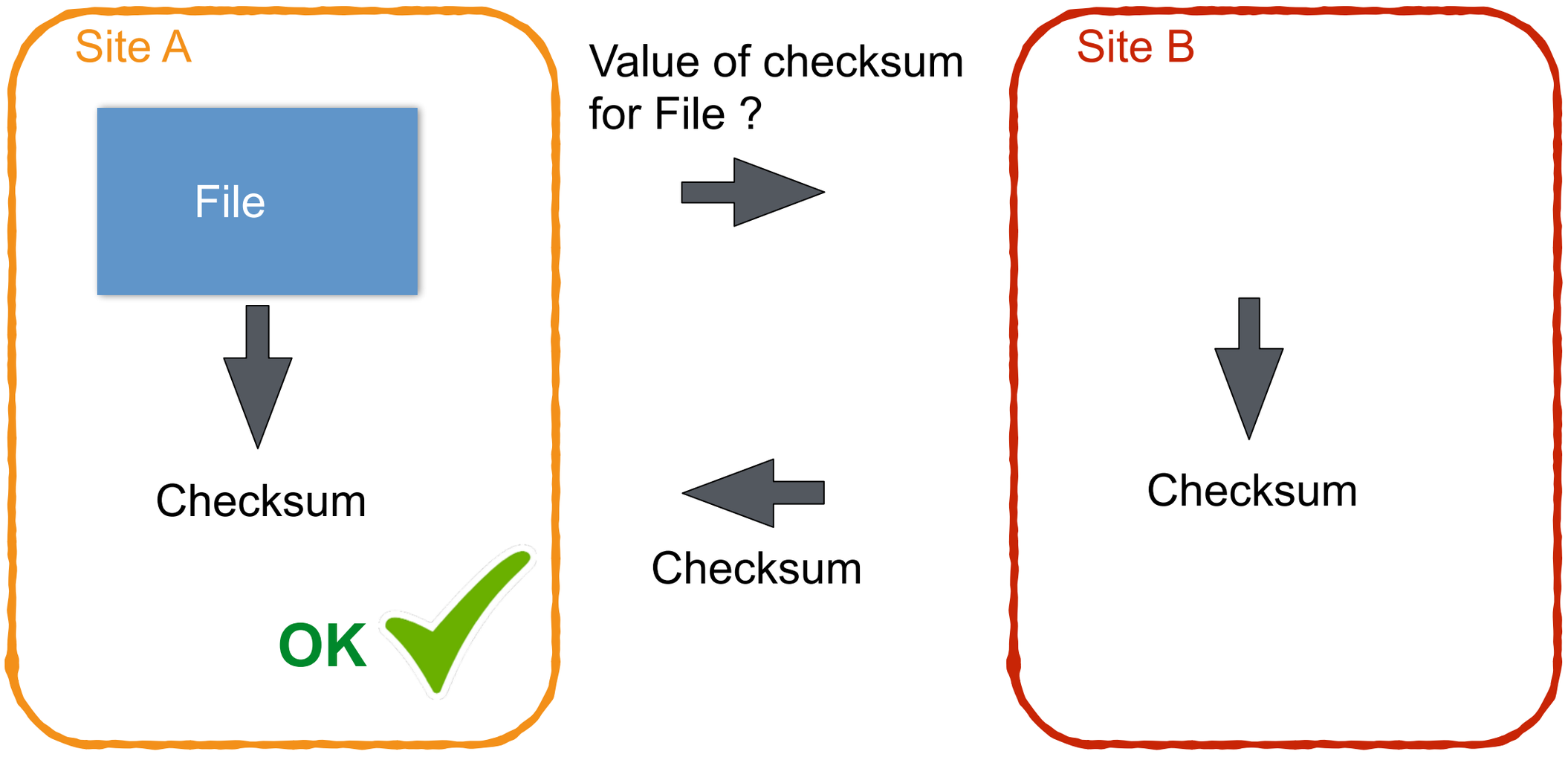
Internal attack : how can we ensure that we really have N copies in the network?



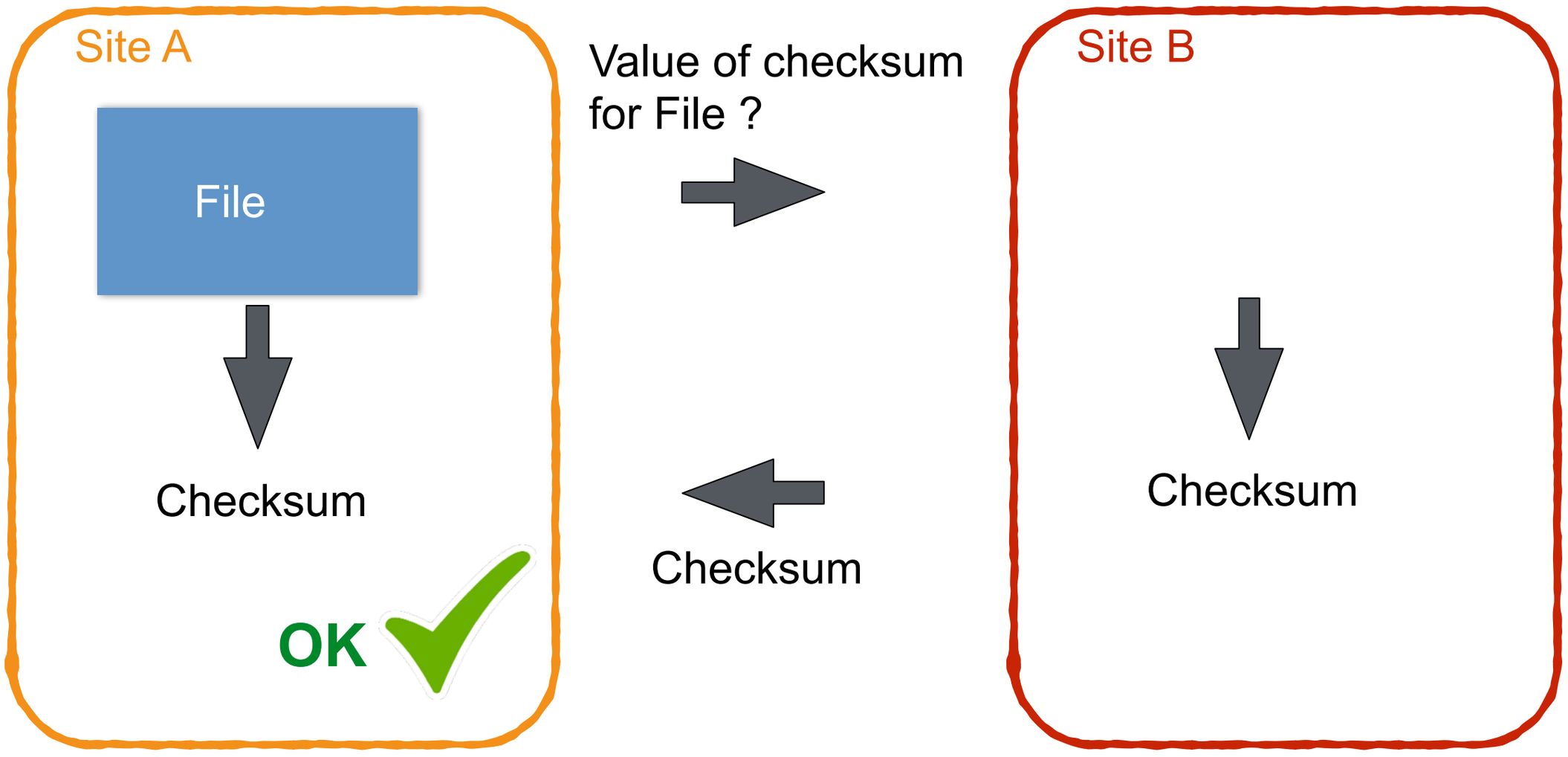
Internal attack : how can we ensure that we really have N copies in the network?



Internal attack : how can we ensure that we really have N copies in the network?



Internal attack : how can we ensure that we really have N copies in the network?



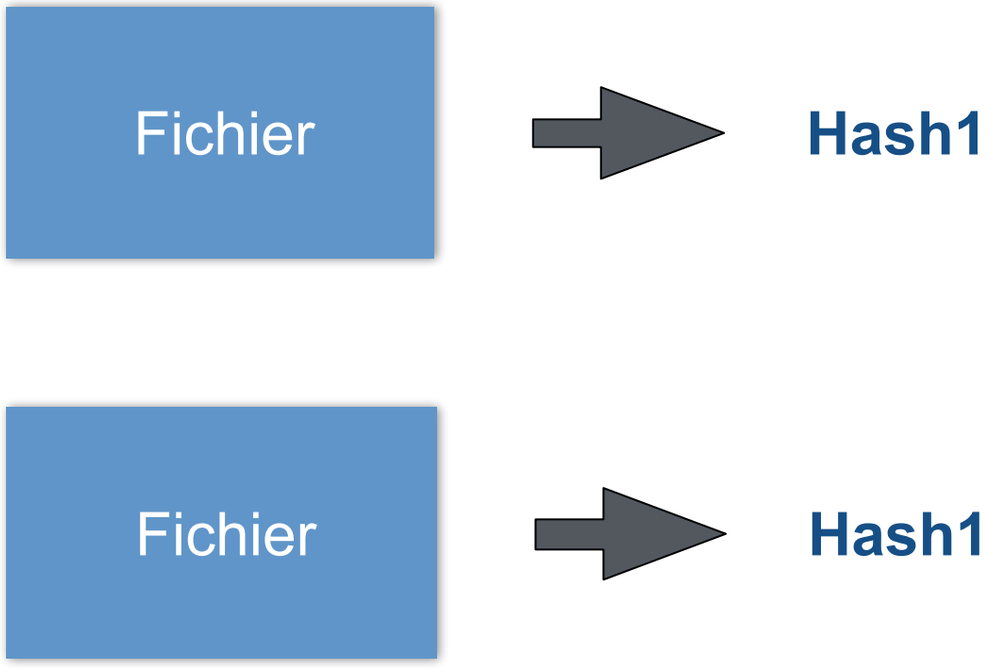
We need to ensure that copies in the network have not been corrupted or deleted.

7. Secure protocol to monitor the integrity of all copies in the network

By requiring to add a one-time-only nonce to the data before producing the checksum, we can ensure that the other party really has a copy of the data and is not just keeping the checksum.

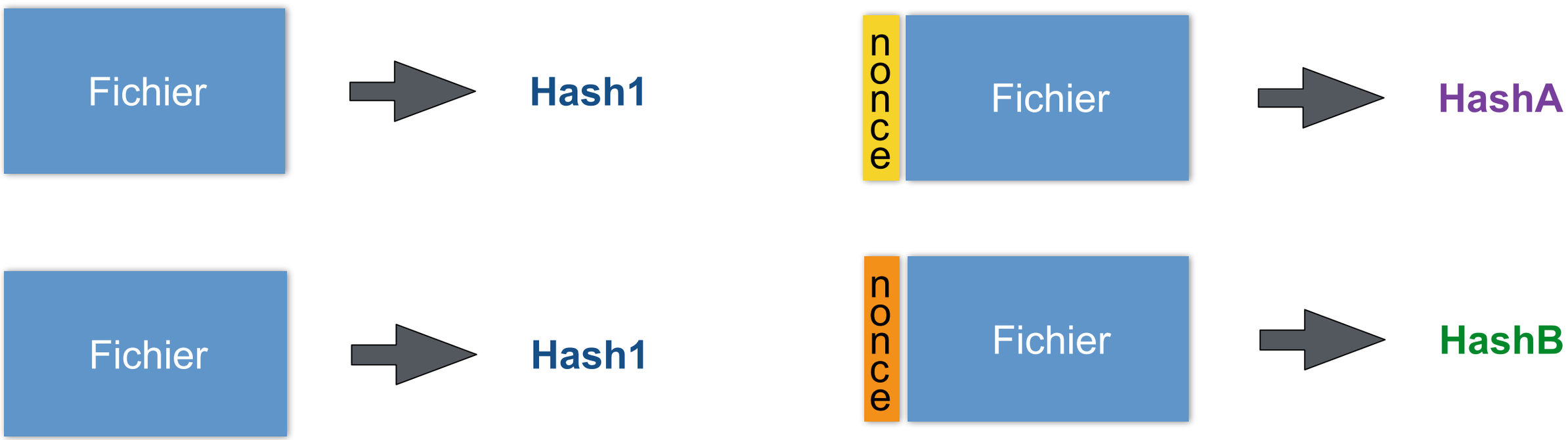
7. Secure protocol to monitor the integrity of all copies in the network

By requiring to add a one-time-only nonce to the data before producing the checksum, we can ensure that the other party really has a copy of the data and is not just keeping the checksum.

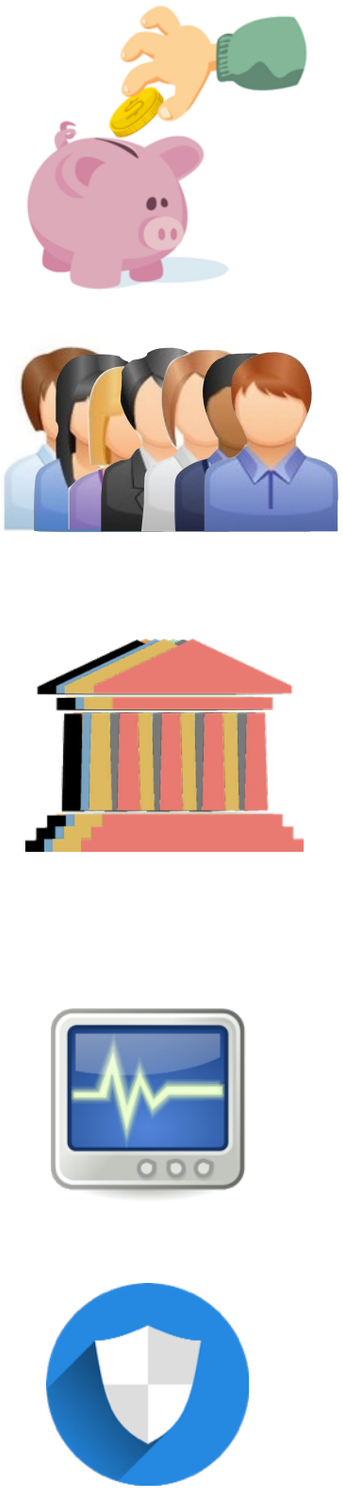


7. Secure protocol to monitor the integrity of all copies in the network

By requiring to add a one-time-only nonce to the data before producing the checksum, we can ensure that the other party really has a copy of the data and is not just keeping the checksum.



7. Secure protocol for the integrity monitoring of all copies



The ideal preservation solution should involve all the aspects previously discussed

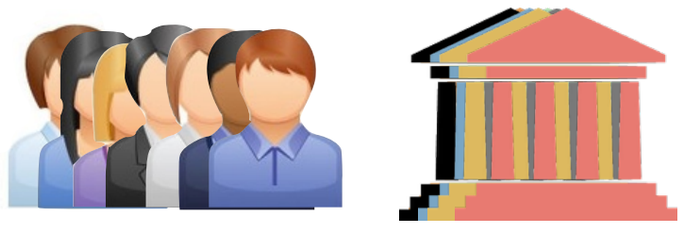
The ideal preservation solution should involve all the aspects previously discussed

Lots of copies
regularly verified
stored on reliable monitored media,
periodically updated,
in a secure software environment,

The ideal preservation solution should involve all the aspects previously discussed

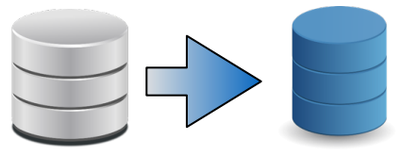
Lots of copies
regularly verified
stored on reliable monitored media,
periodically updated,
in a secure software environment,

managed by different people
in independent institutions



The ideal preservation solution should involve all the aspects previously discussed

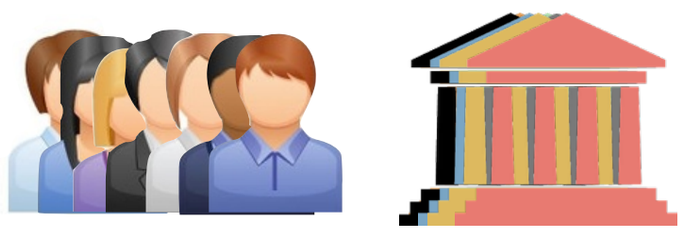
Lots of copies
regularly verified
stored on reliable monitored media,
periodically updated,
in a secure software environment,



managed by different people
in independent institutions

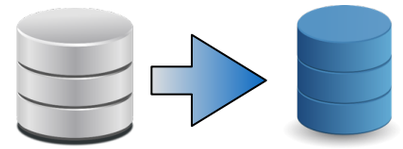
and at low cost.

The ideal preservation solution should involve all the aspects previously discussed



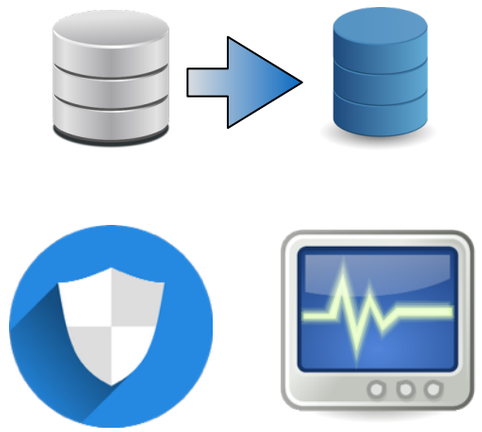
The ideal preservation solution should involve all the aspects previously discussed

Technology

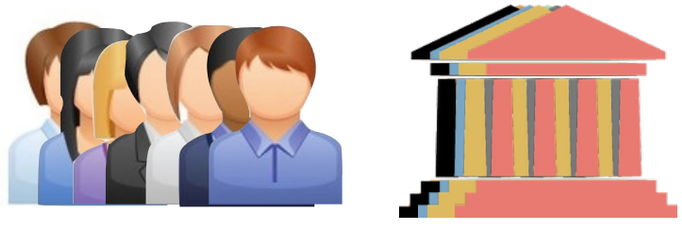


The ideal preservation solution should involve all the aspects previously discussed

Technology

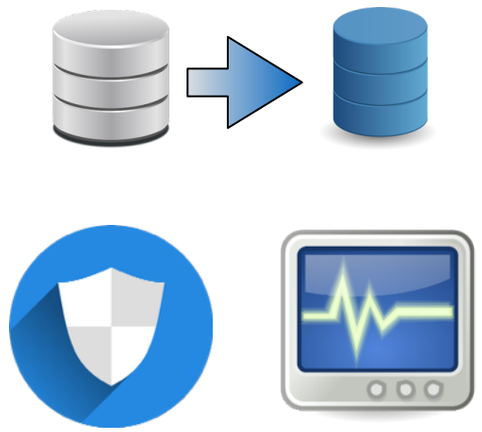


Organization

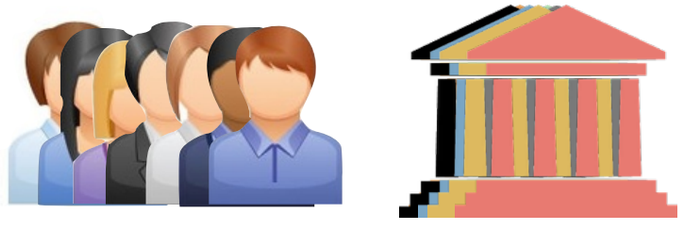


The ideal preservation solution should involve all the aspects previously discussed

Technology

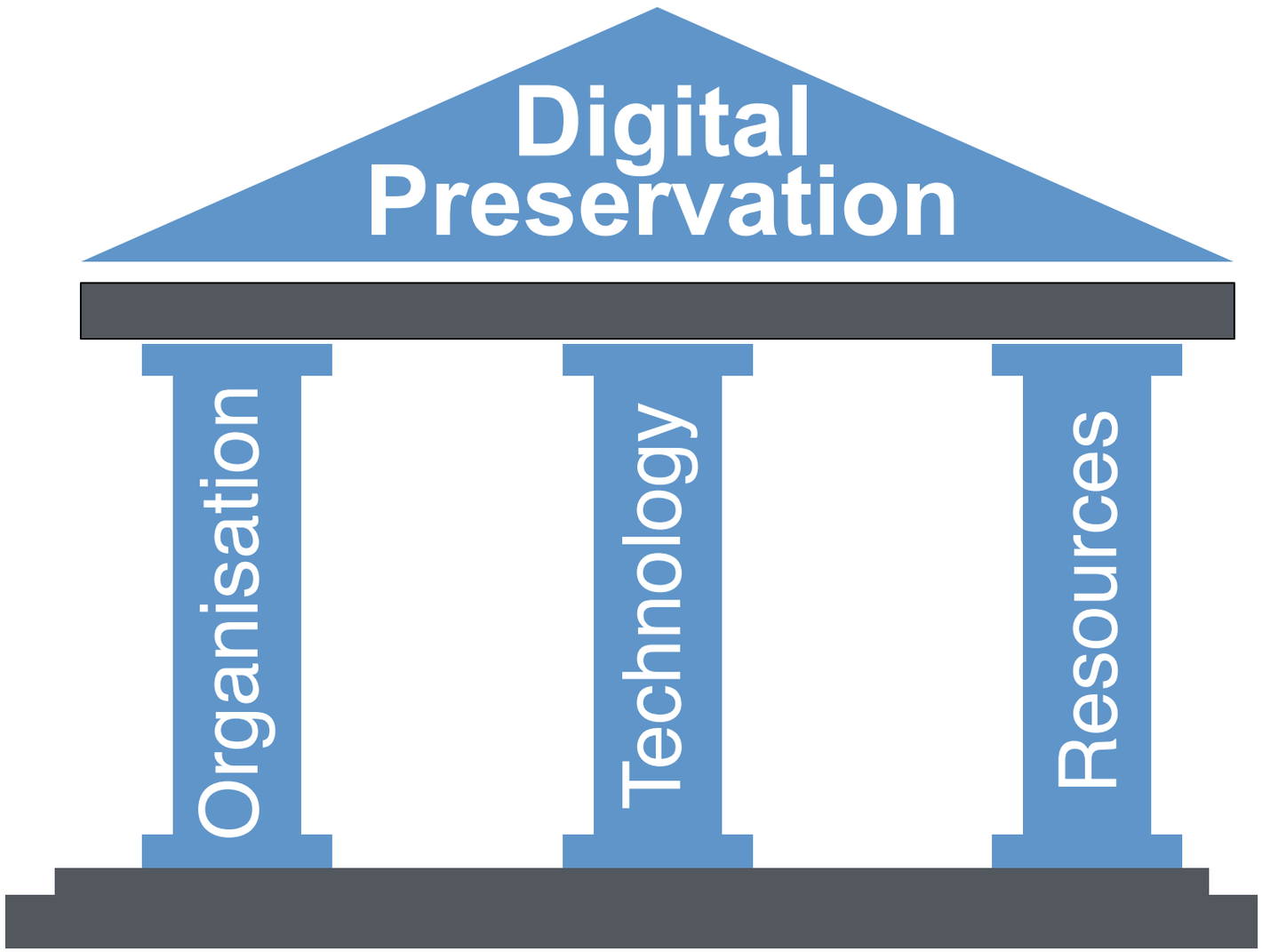


Organization



Resources

The 3 pillars of digital preservation: organisation, ressources and technology



There is a whole spectrum of preservation grades

When is a digital preservation solution good **enough** ?

There is a whole spectrum of preservation grades

When is a digital preservation solution good **enough** ?

It all depends on your threat-model (which includes your limited resources).

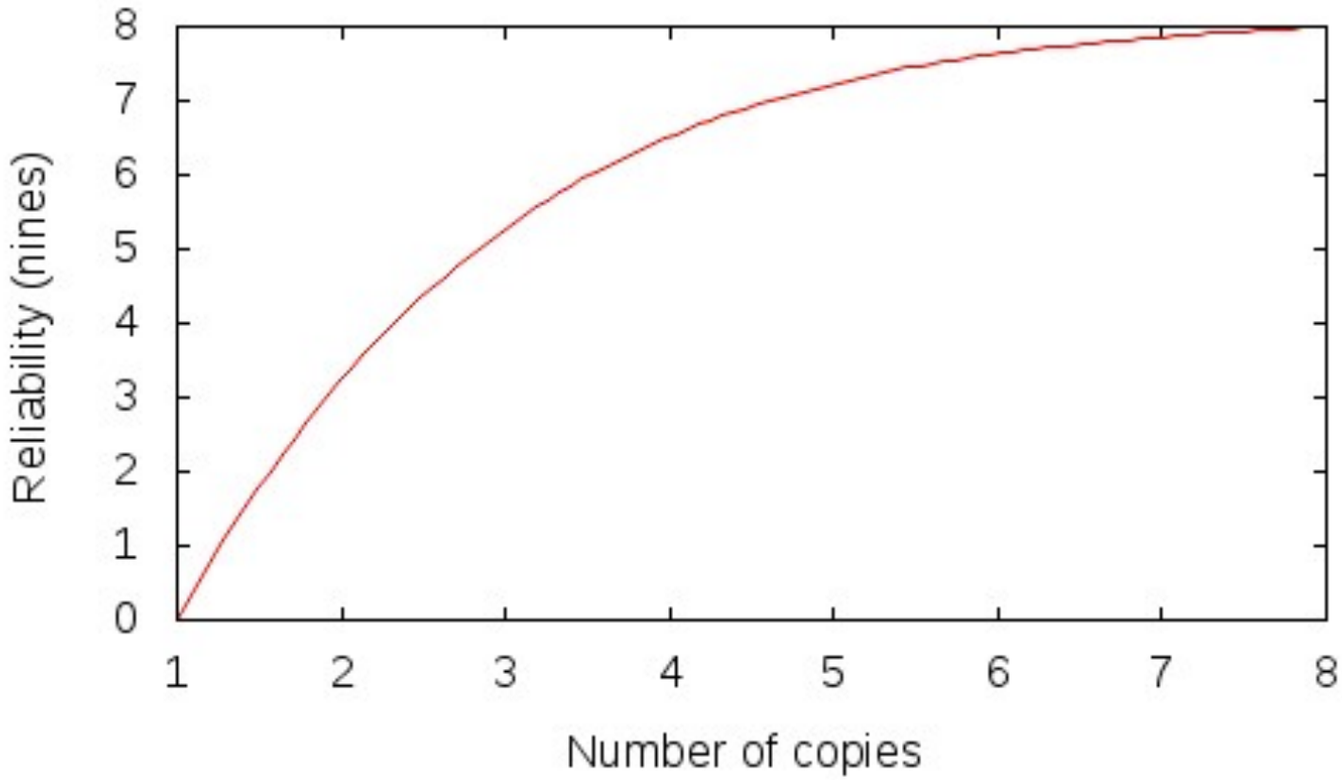
How many copies **?**

Ideally, it should be evaluated based on a detailed **risk and threat assessment**.

https://blog.dshr.org/2011/03/how-few-copies.html

How **many** copies **?**

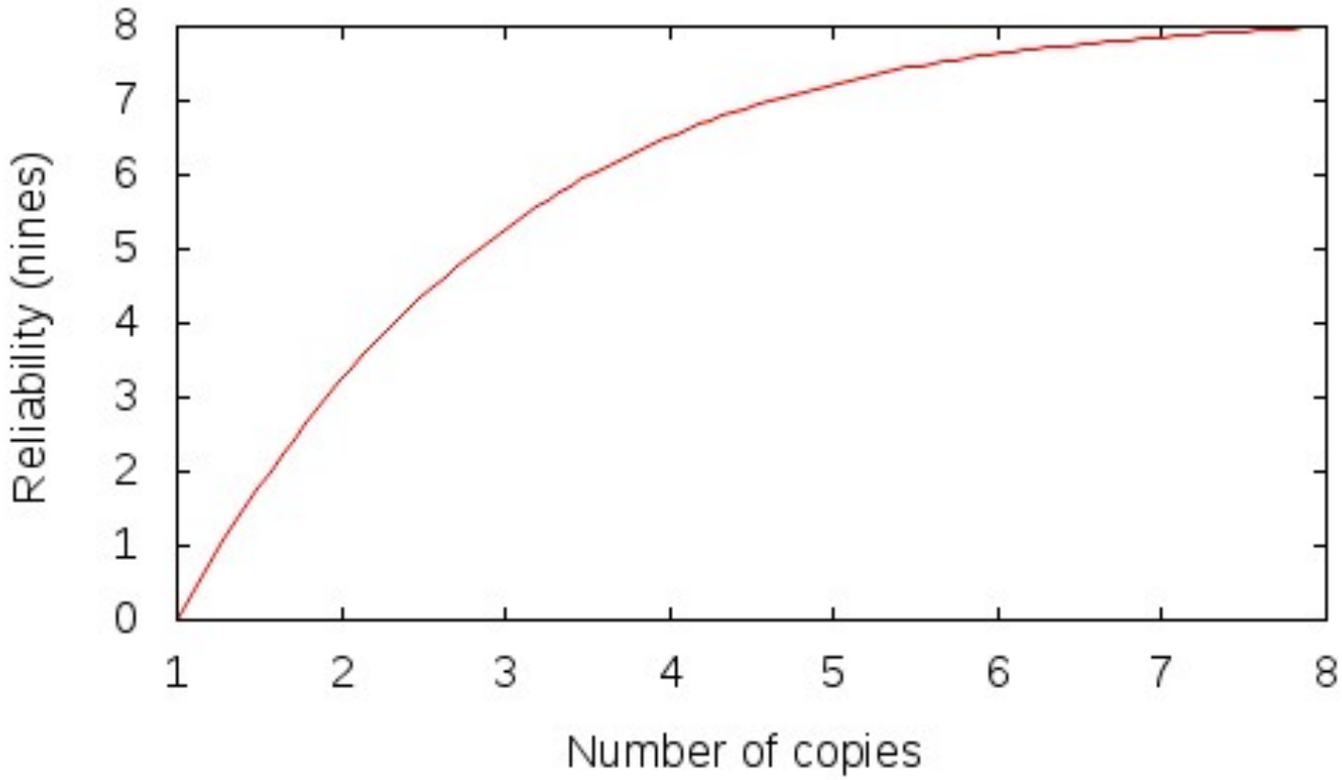
Ideally, it should be evaluated based on a detailed **risk and threat assessment**.



https://blog.dshr.org/2011/03/how-few-copies.html

How **many** copies **?**

Ideally, it should be evaluated based on a detailed **risk and threat assessment**.

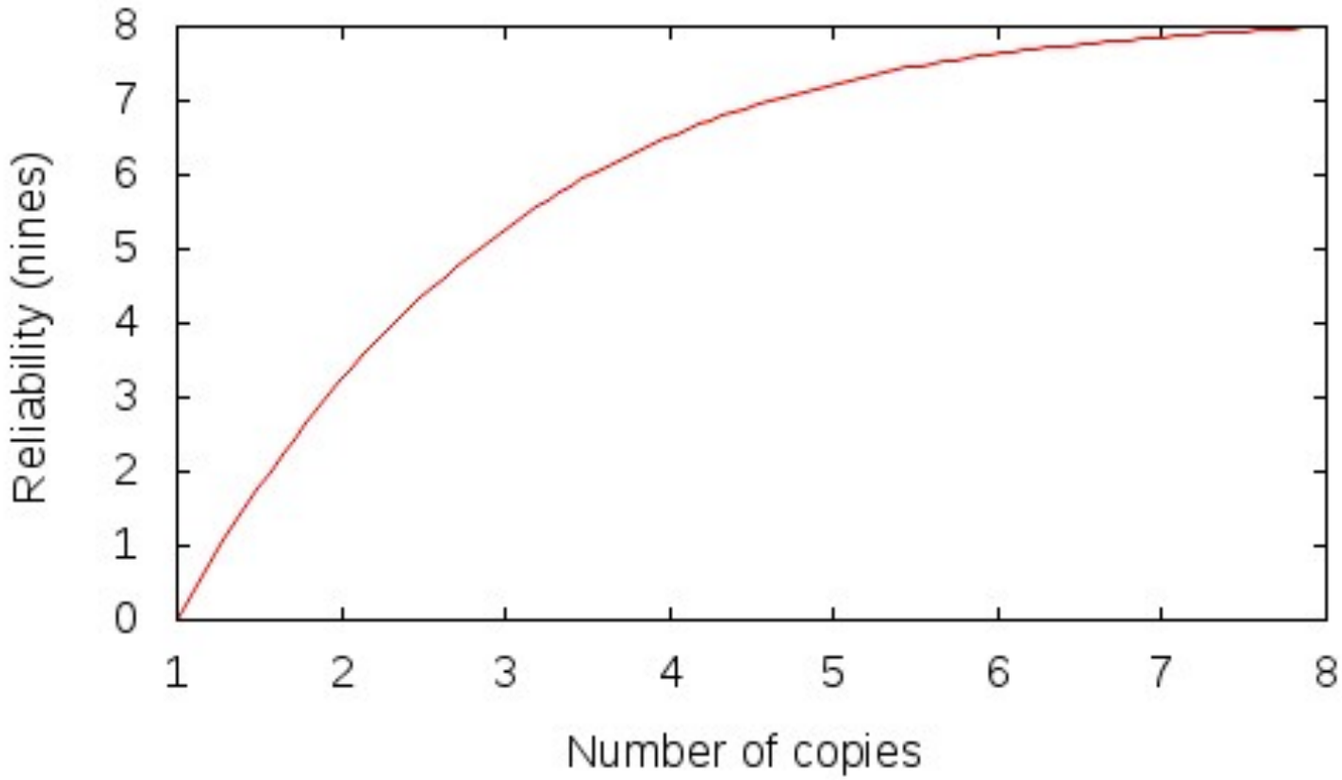


https://blog.dshr.org/2011/03/how-few-copies.html

But in practice, the reliability is very difficult to evaluate.

How **many** copies **?**

Ideally, it should be evaluated based on a detailed **risk and threat assessment**.



https://blog.dshr.org/2011/03/how-few-copies.html

But in practice, the reliability is very difficult to evaluate.

Hence, the more copies, the safer

Lots of **C**opies **K**eep **S**tuff **S**afe

Lots of **C**opies **K**eep **S**tuff **S**afe

Lots of Copies Keep Stuff Safe

Less aggressive compression

Lots of **C**opies **K**eep **S**tuff **S**afe

Faster failures detection and repair

Less aggressive compression

Lots of Copies Keep Stuff Safe

More reliable copies

Faster failures detection and repair

Less aggressive compression

Lots of Copies Keep Stuff Safe

Less correlated copies

More reliable copies

Faster failures detection and repair

Less aggressive compression

Lots of Copies Keep Stuff Safe

More copies

Less correlated copies

More reliable copies

Faster failures detection and repair

Less aggressive compression

Let’s have a break

Agenda

Motivation

LOCKSS main concepts

SAFE LOCKSS network : a Case Study

Build your own LOCKSS Network

Discussions

LOCKSS is, first of all, a **technology**

open source software

originally for PCA (Global LOCKSS Network)

awarded technology (perfect score by CRL)

a robust integrity check and repair protocol

And a **community** :

a large variety of networks based on LOCKSS technology

And a **community** :

a large variety of networks based on LOCKSS technology

CLOCKSS

Cariniana

Meta Archive

SAFE

PKP
PUBLIC KNOWLEDGE PROJECT

FDSYS
FEDERAL DIGITAL SYSTEM

LOCKSS

DIGITAL COMMONS
powered by bepress

Perma.cc

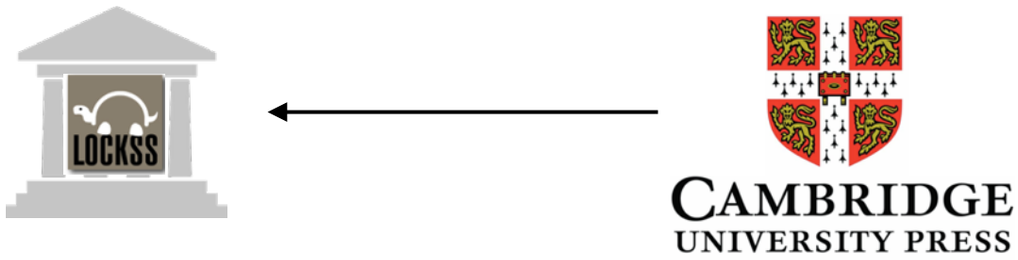
WestVault

COPPUL

THE ALABAMA DIGITAL PRESERVATION NETWORK
PRESERVING ALABAMA'S DIGITAL RESOURCES

LEARNING RESOURCES

How does LOCKSS work in practice ?
Illustration with the **Global LOCKSS Network**



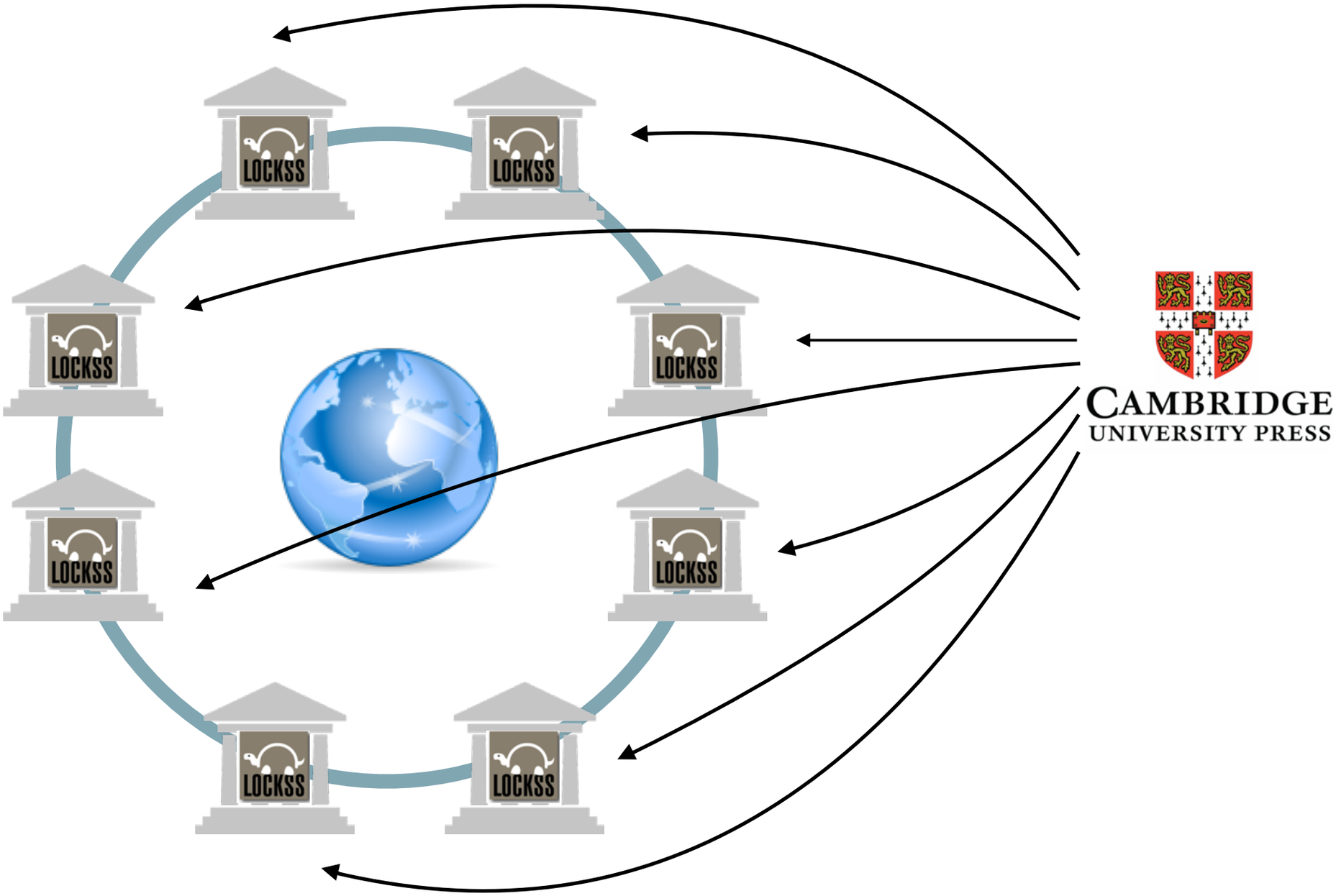
LEARNING RESOURCES

How does LOCKSS work in practice ?
Illustration with the **Global LOCKSS Network**

CAMBRIDGE
UNIVERSITY PRESS

LEARNING RESOURCES

How does LOCKSS work in practice ?
Illustration with the **Global LOCKSS Network**



LEARNING RESOURCES

How does LOCKSS work in practice ?
Illustration with the **Global LOCKSS Network**

The same software can be used to preserve any resource

http://difusion.academiewb.be/

LOCKSS in practice: ingestion of content

Administrative server

network configuration
[XML]
global parameters

title database
[XML]
Plug-in location
AUs definitions
list of caches IP address

plug-ins repository
CVS for JARs [java | xml]
IR1 certified plug-in
IR2 certified plug-in
IR3 certified plug-in

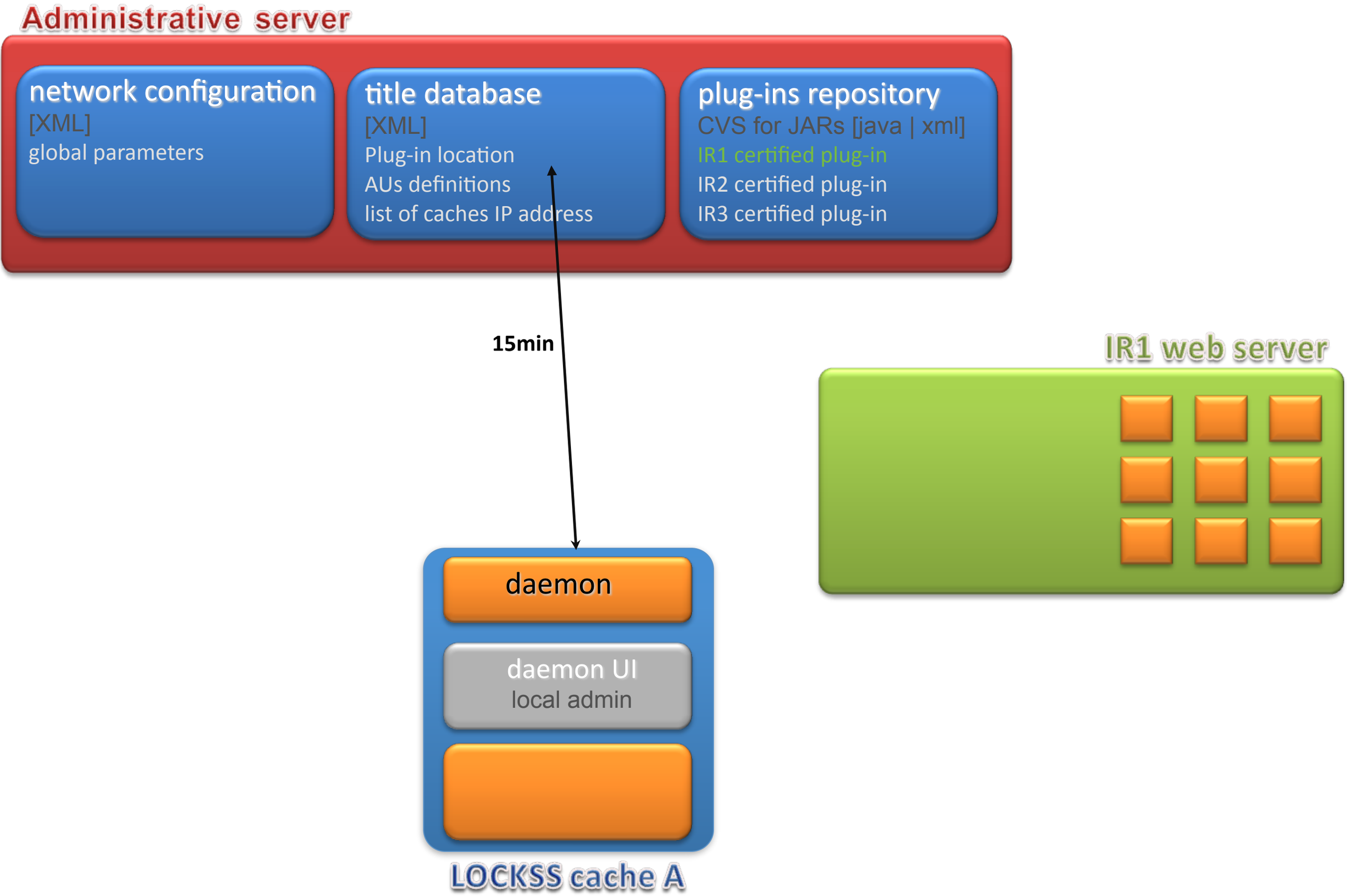
IR1 web server

daemon

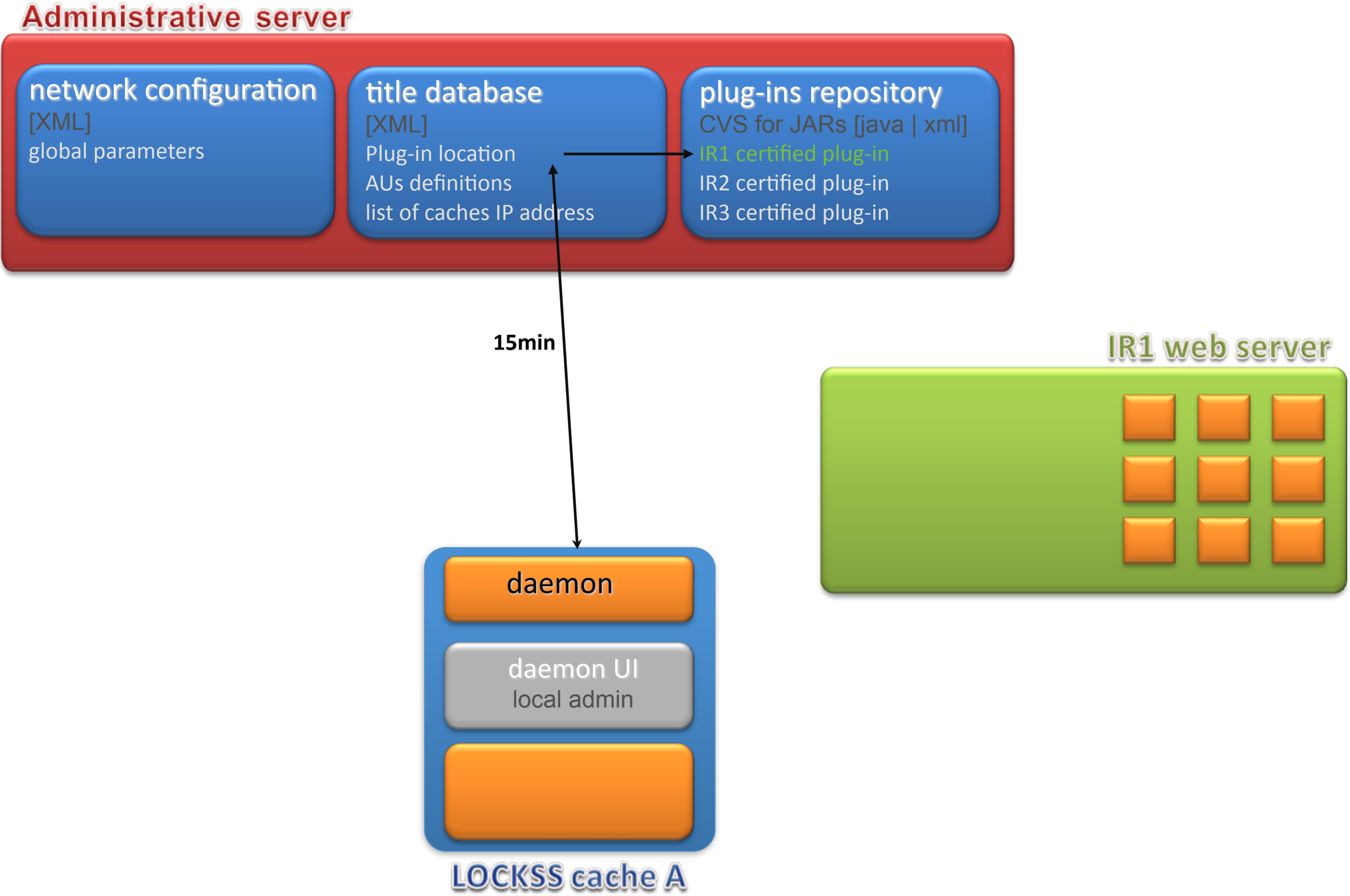
daemon UI
local admin

LOCKSS cache A

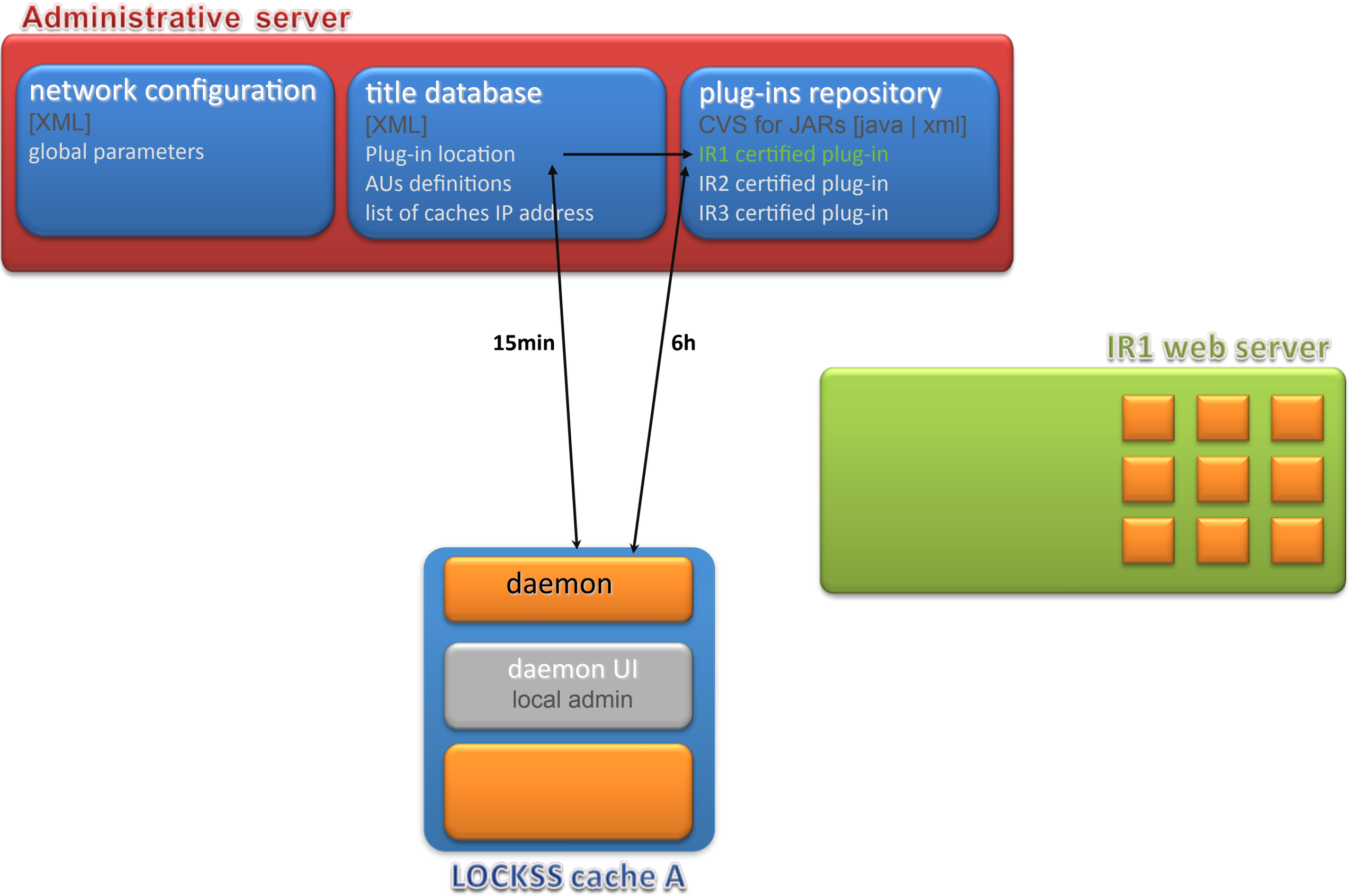
LOCKSS in practice: ingestion of content



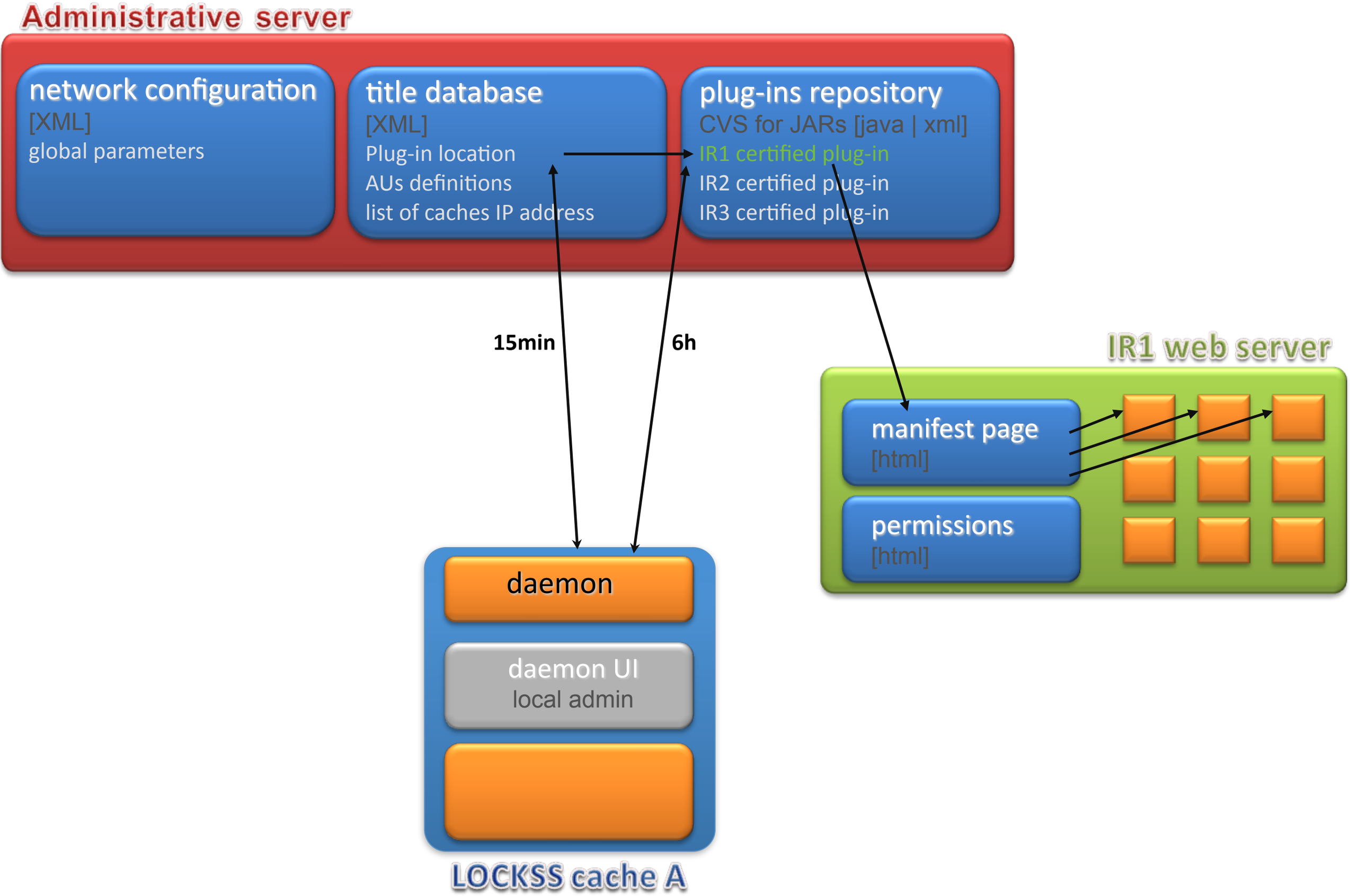
LOCKSS in practice: ingestion of content



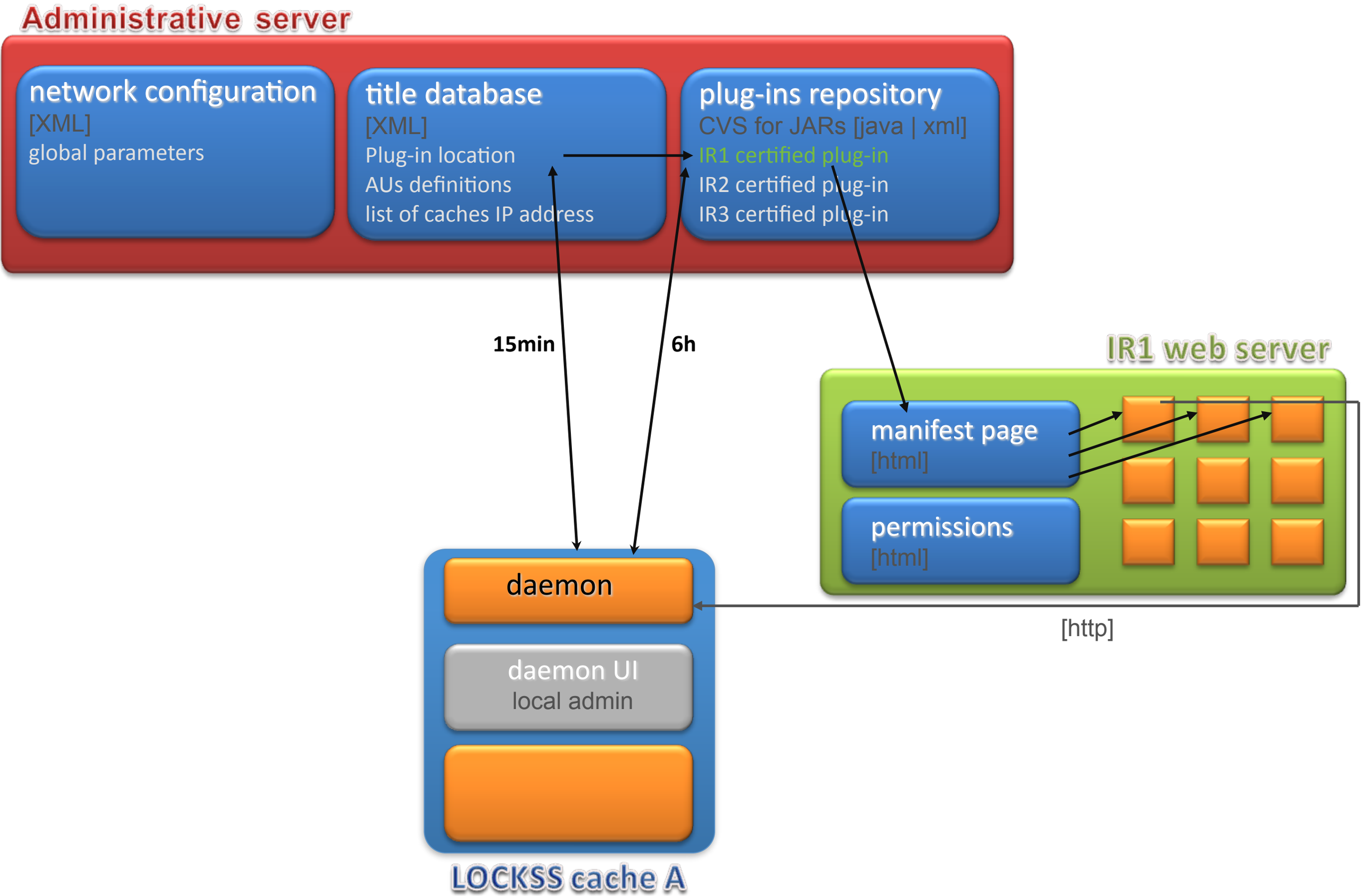
LOCKSS in practice: ingestion of content



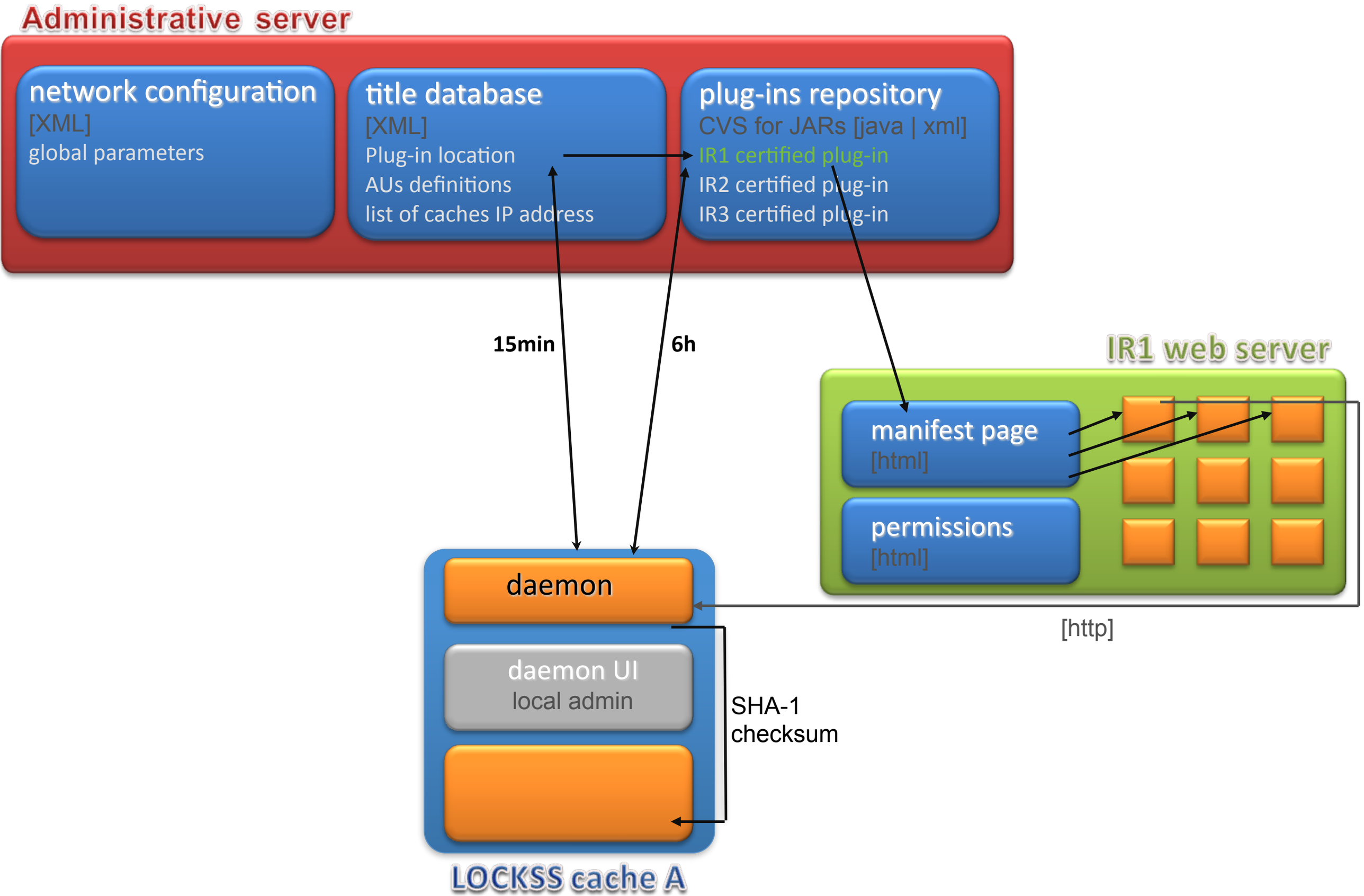
LOCKSS in practice: ingestion of content



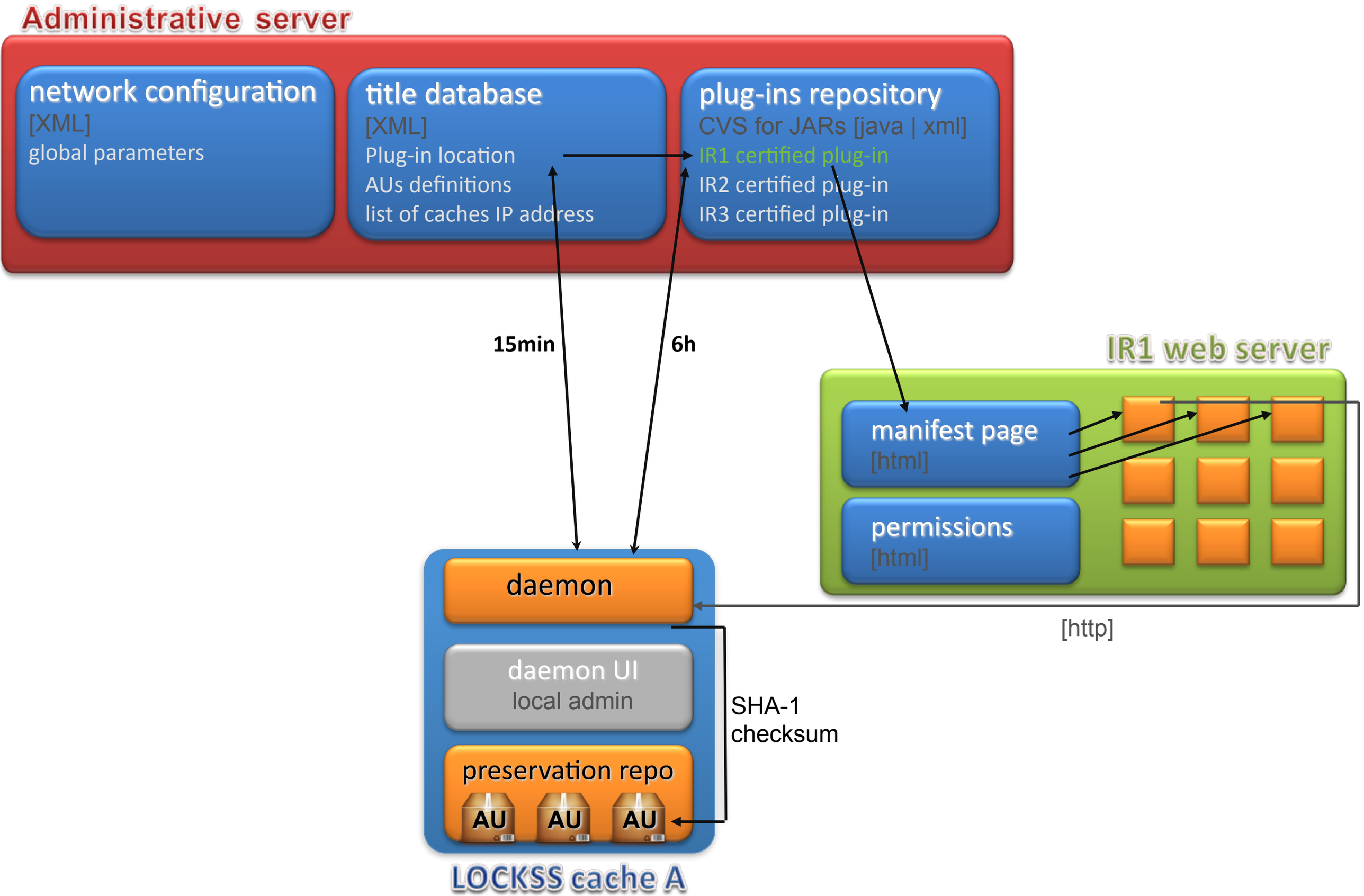
LOCKSS in practice: ingestion of content



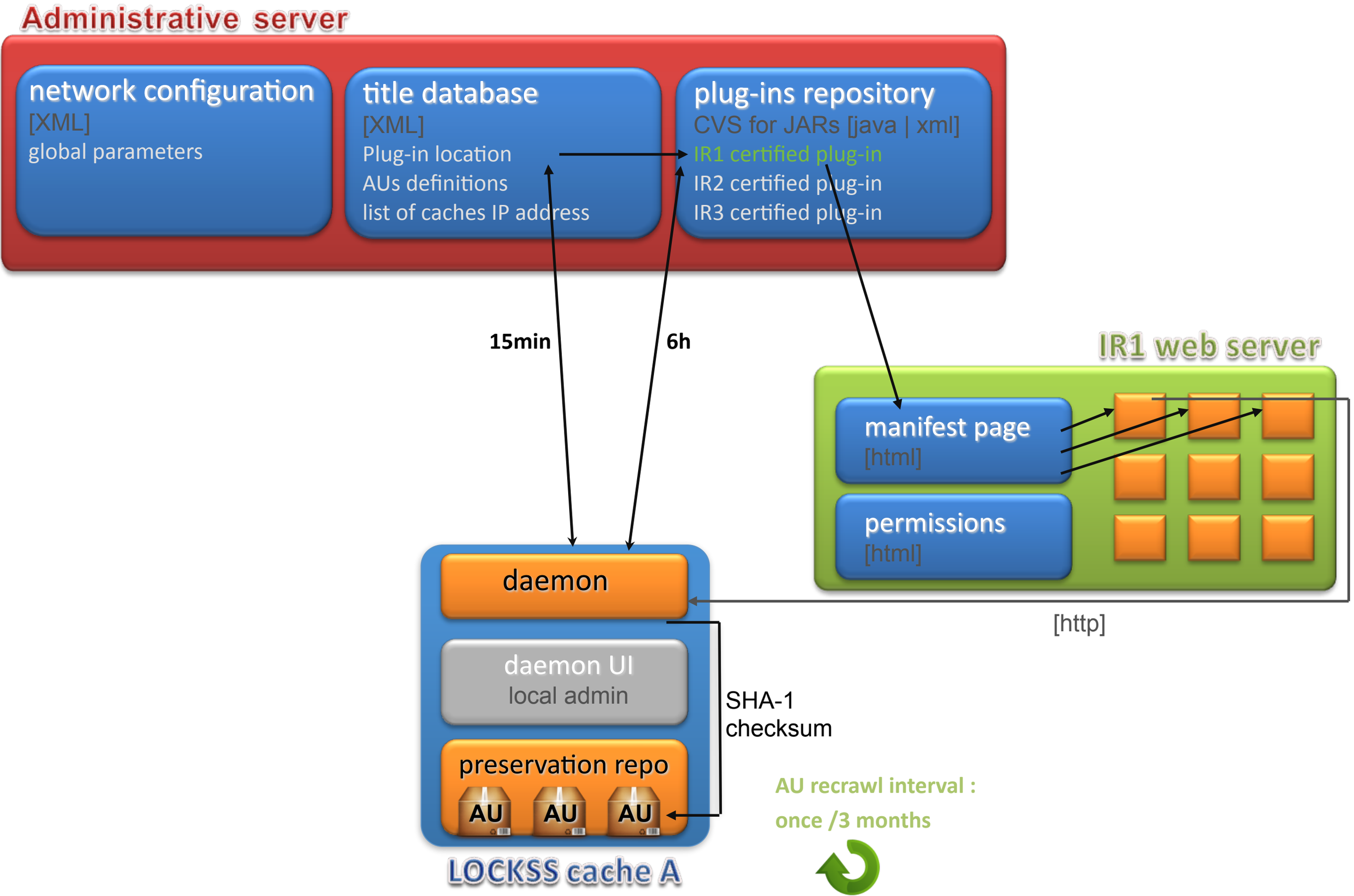
LOCKSS in practice: ingestion of content



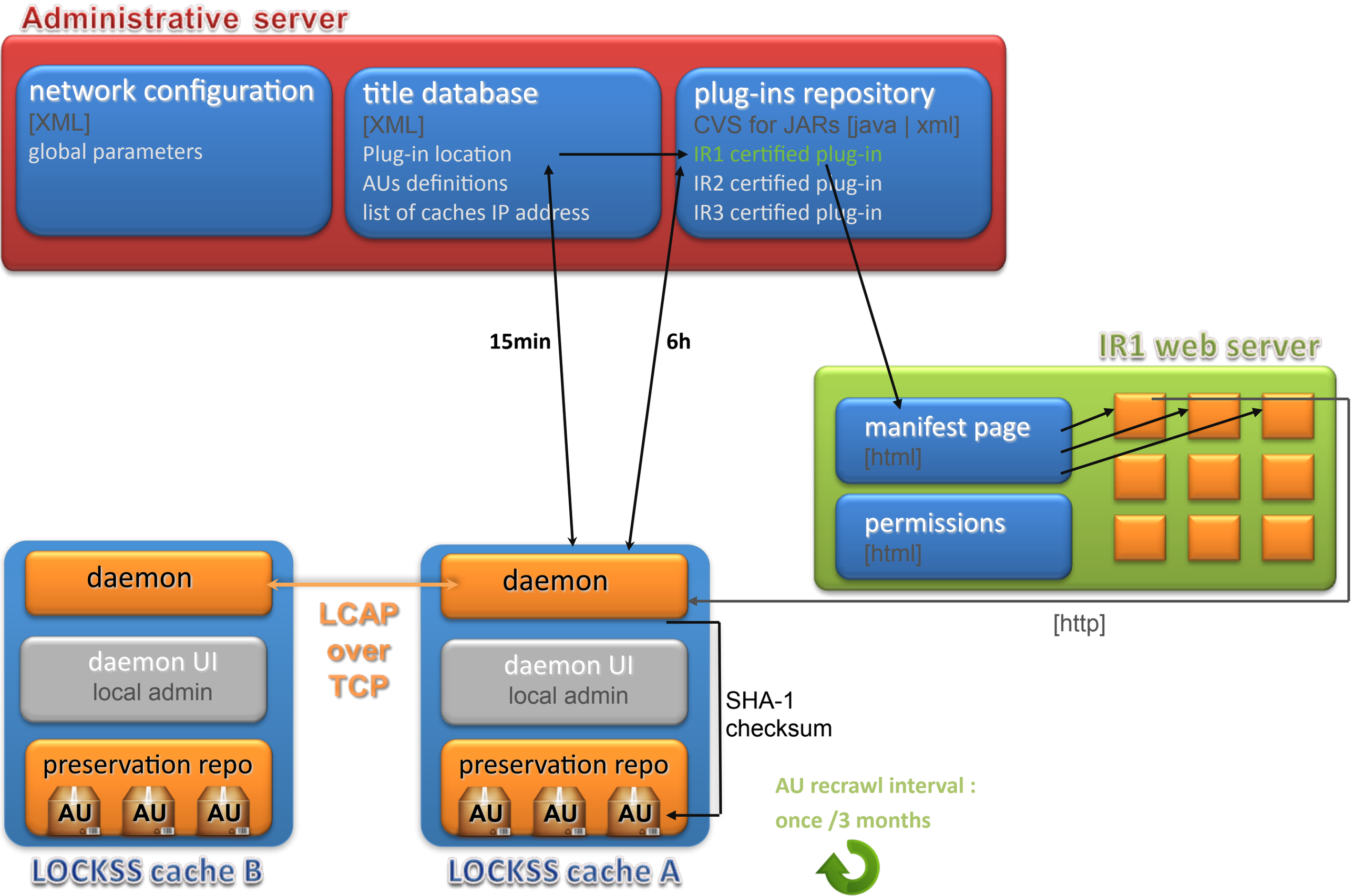
LOCKSS in practice: ingestion of content



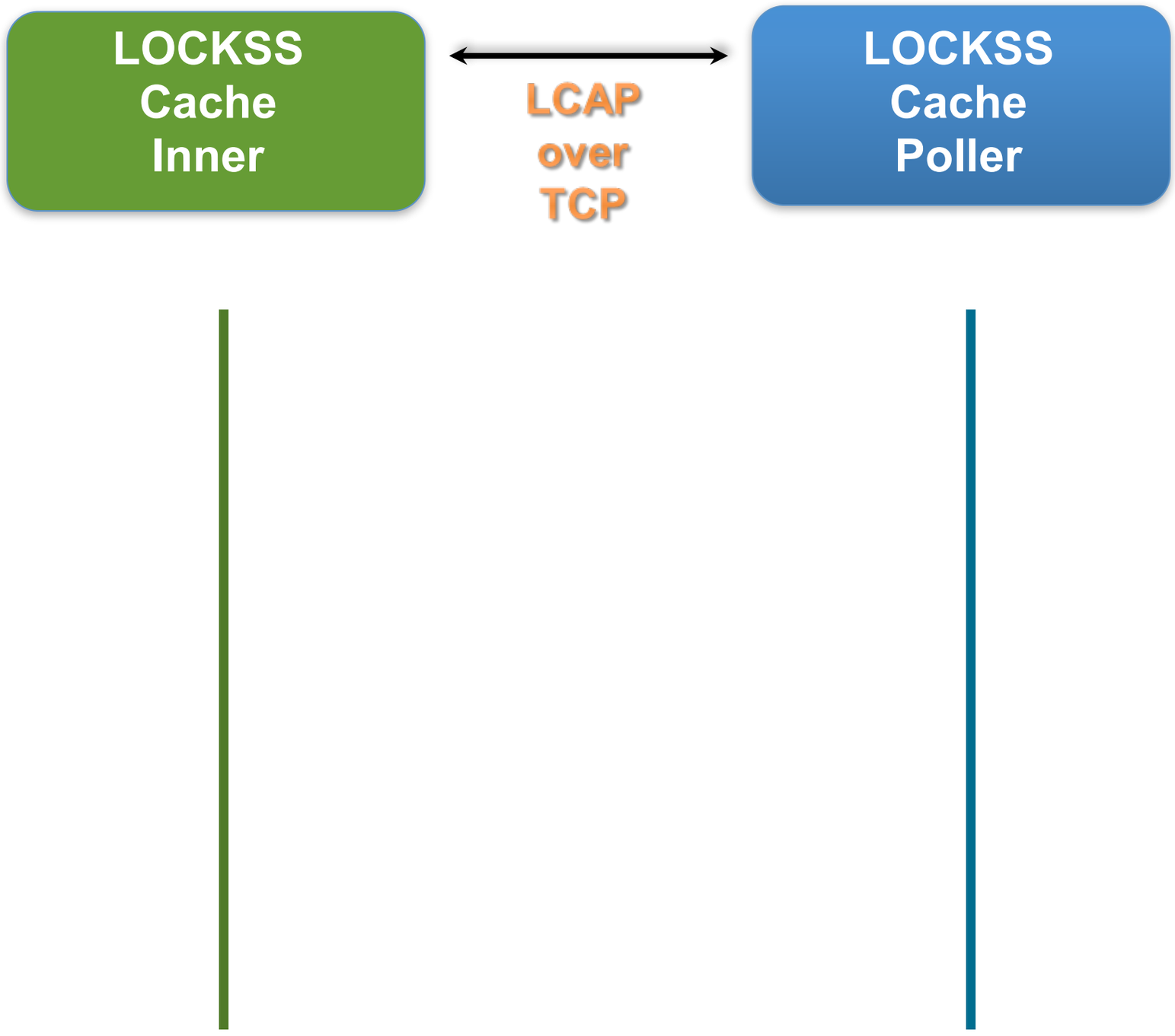
LOCKSS in practice: ingestion of content



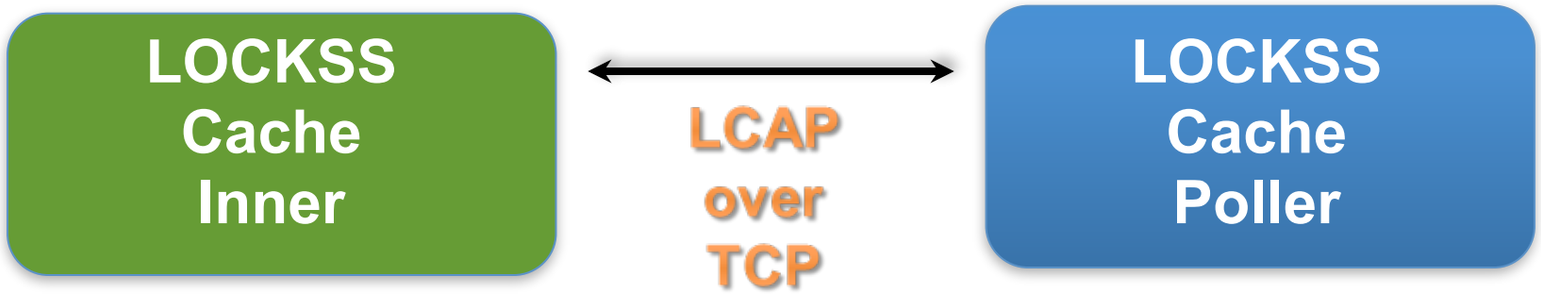
LOCKSS in practice: ingestion of content

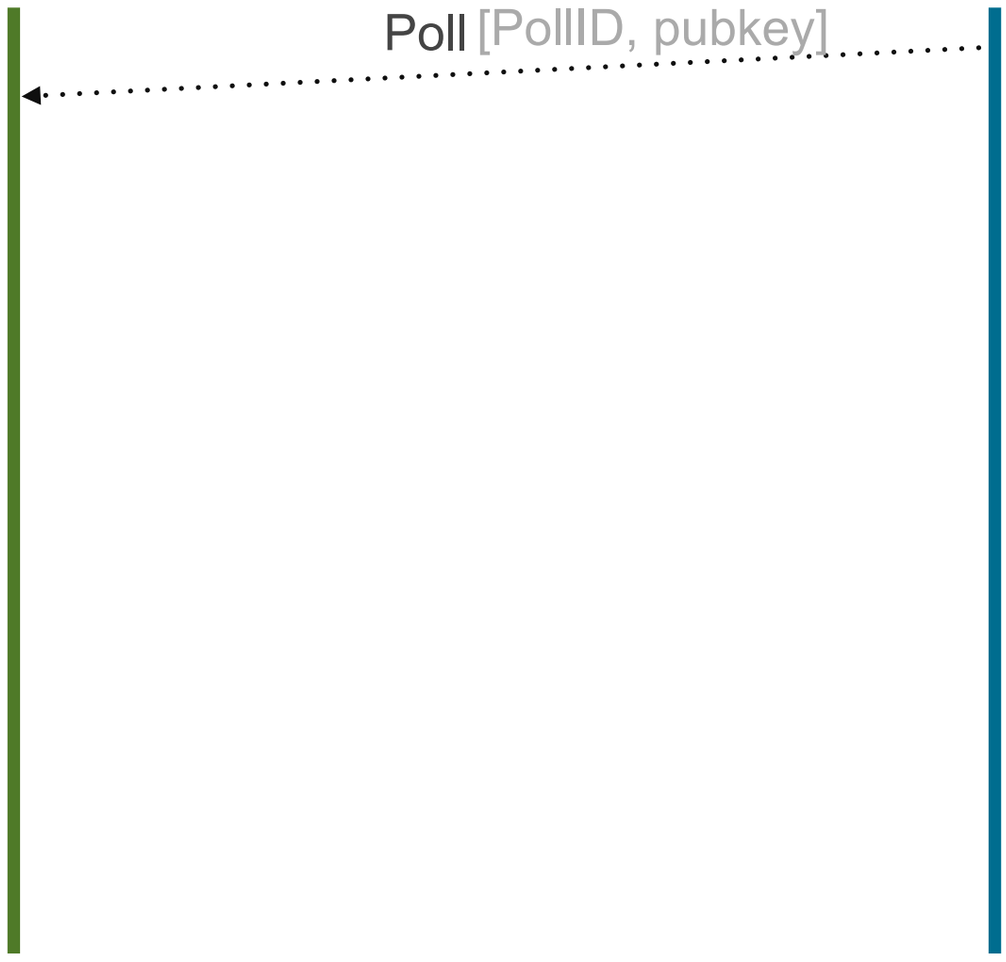


LOCKSS in practice: cache integrity monitoring

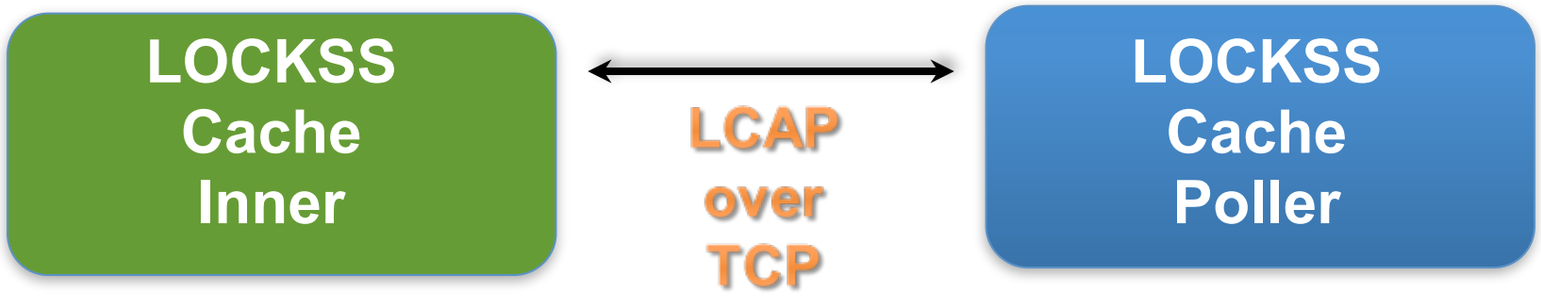


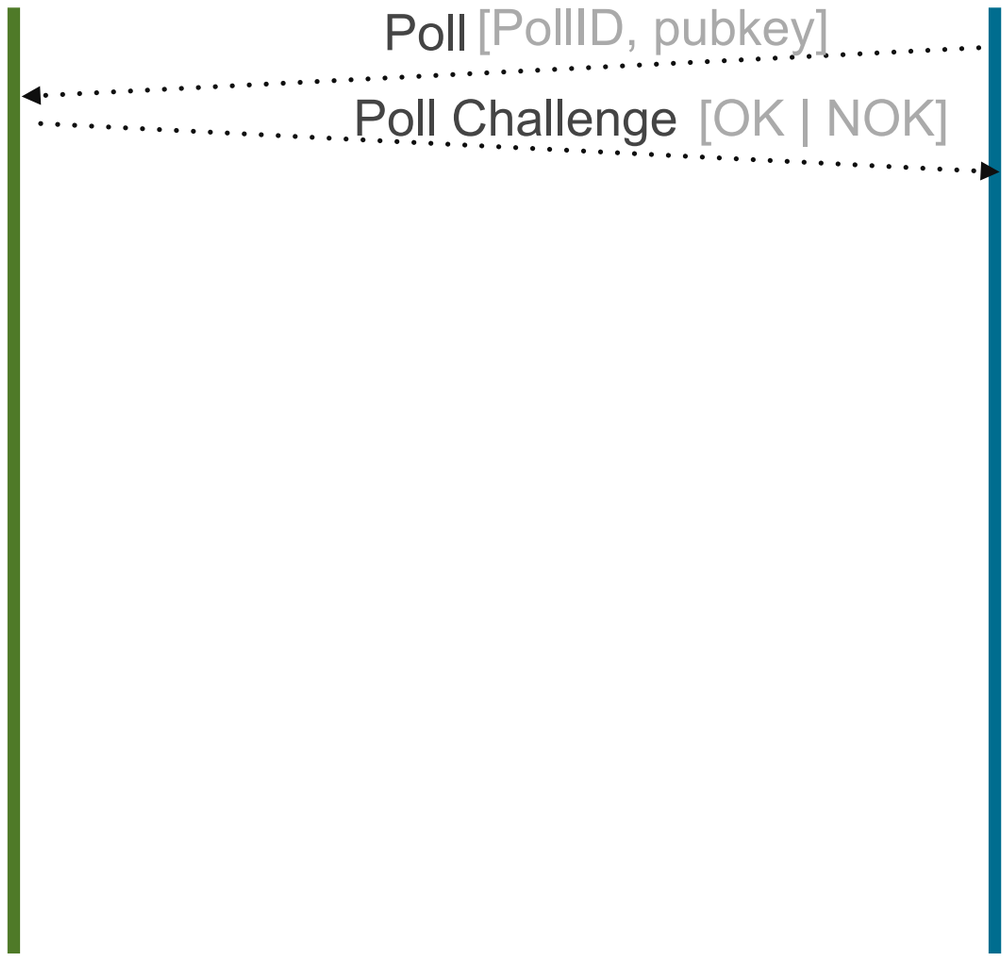
LOCKSS in practice: cache integrity monitoring



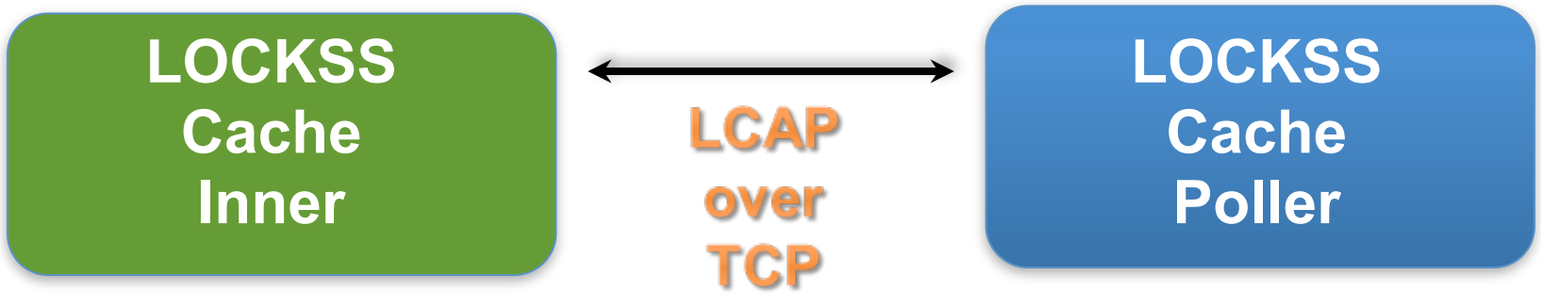


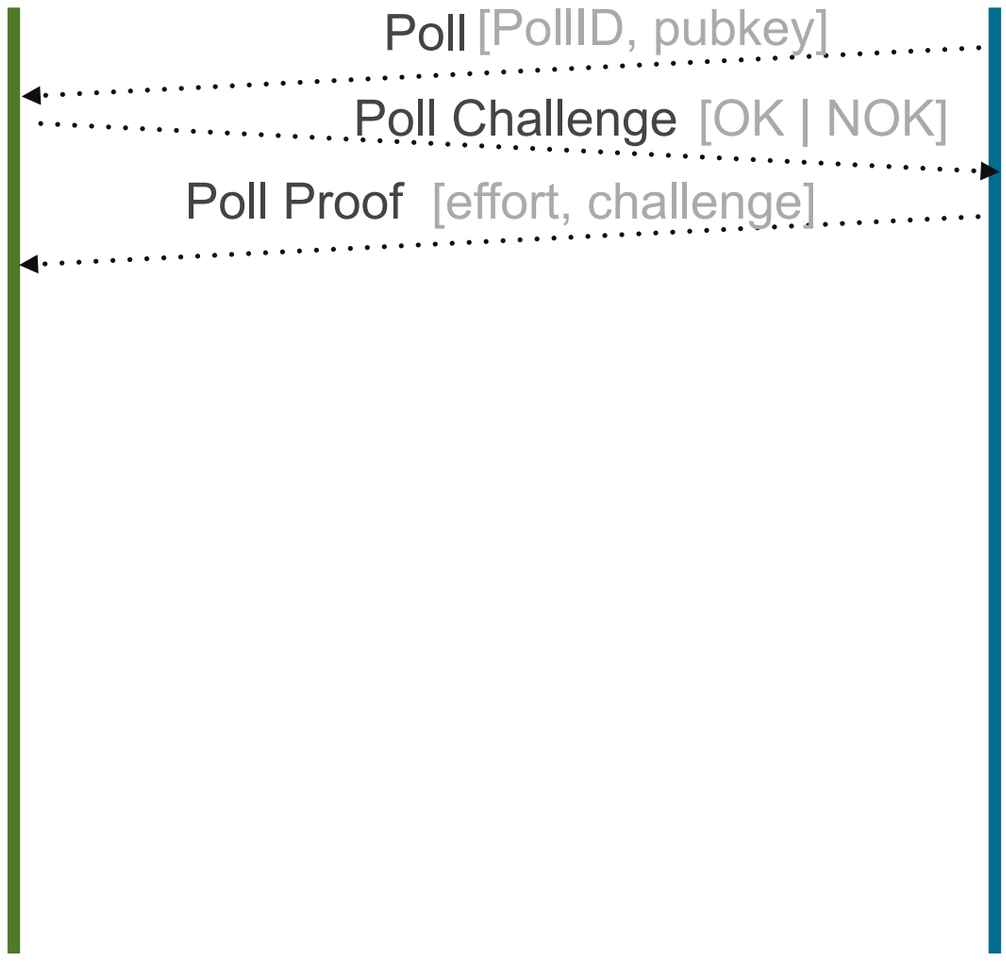
LOCKSS in practice: cache integrity monitoring



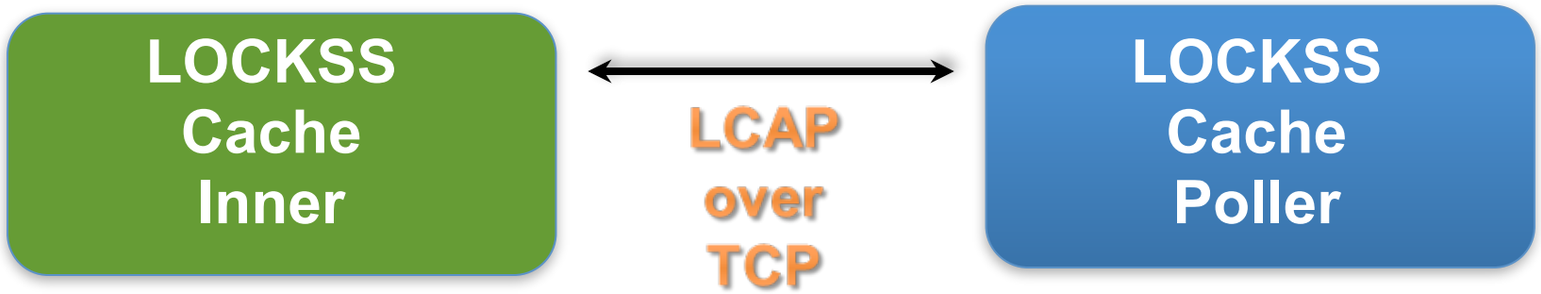


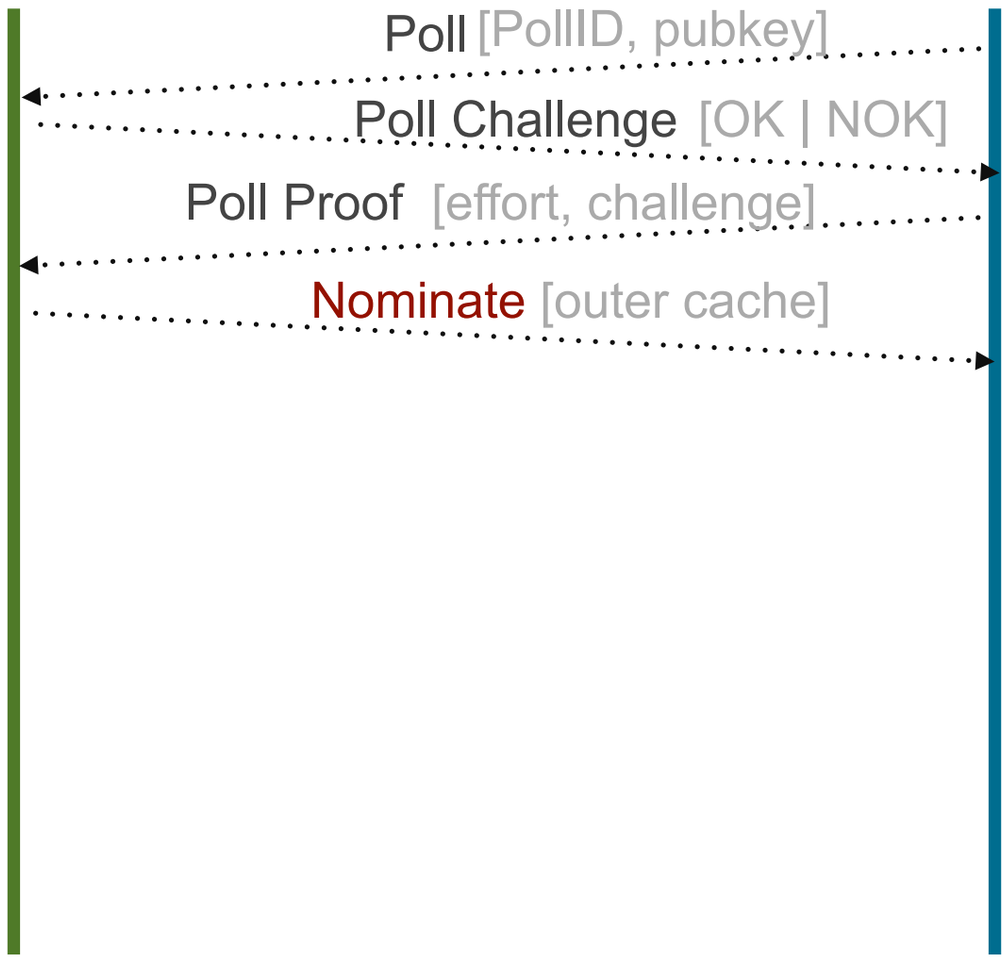
LOCKSS in practice: cache integrity monitoring



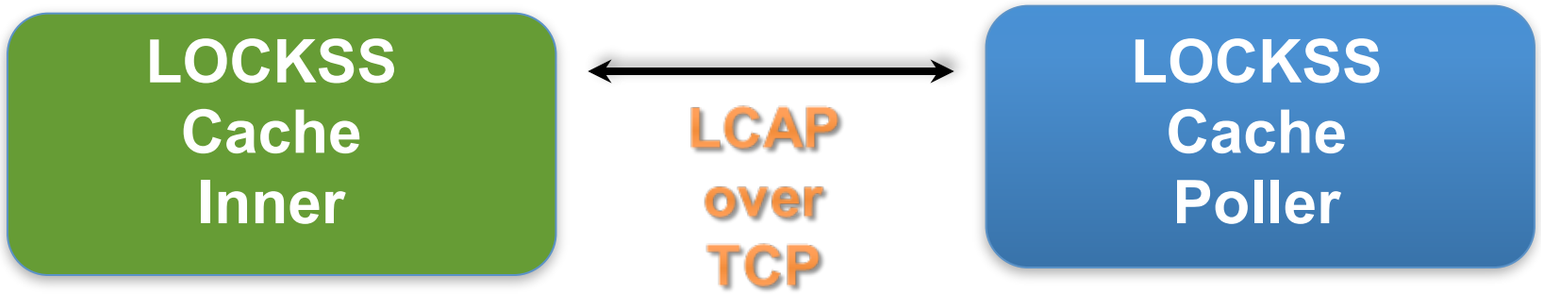


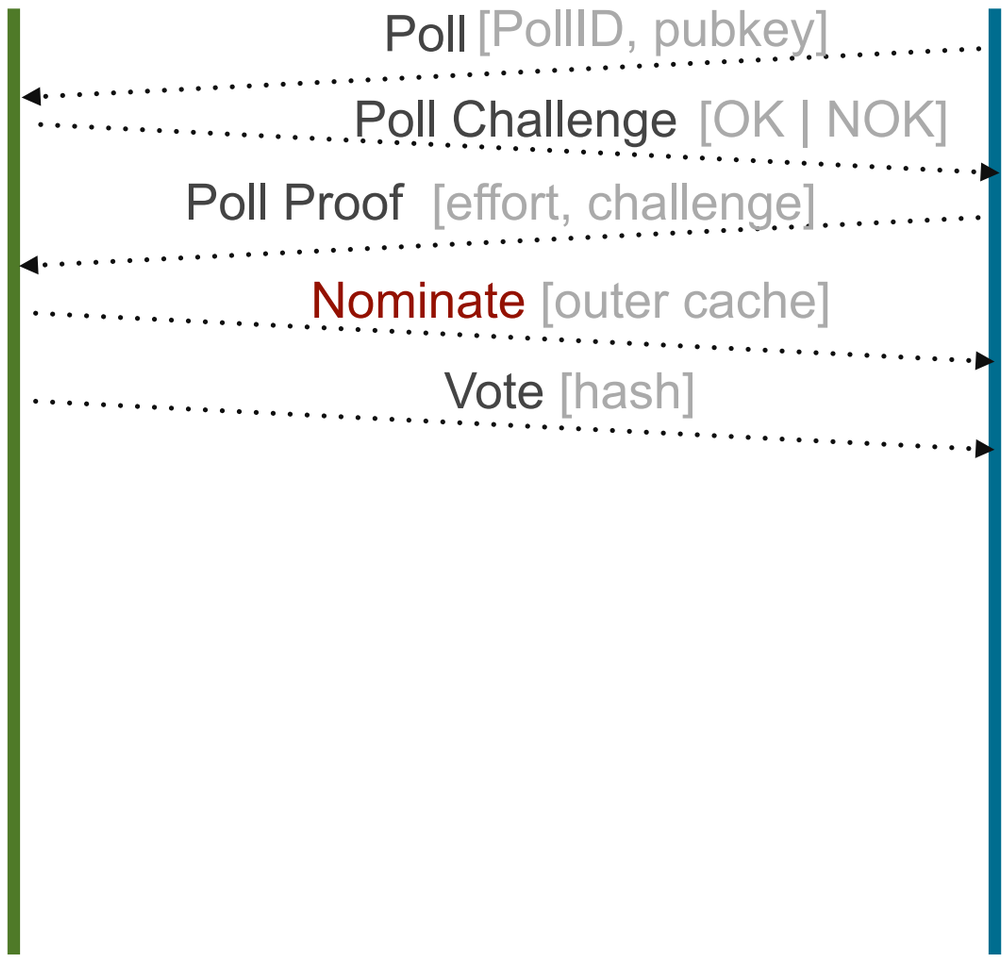
LOCKSS in practice: cache integrity monitoring



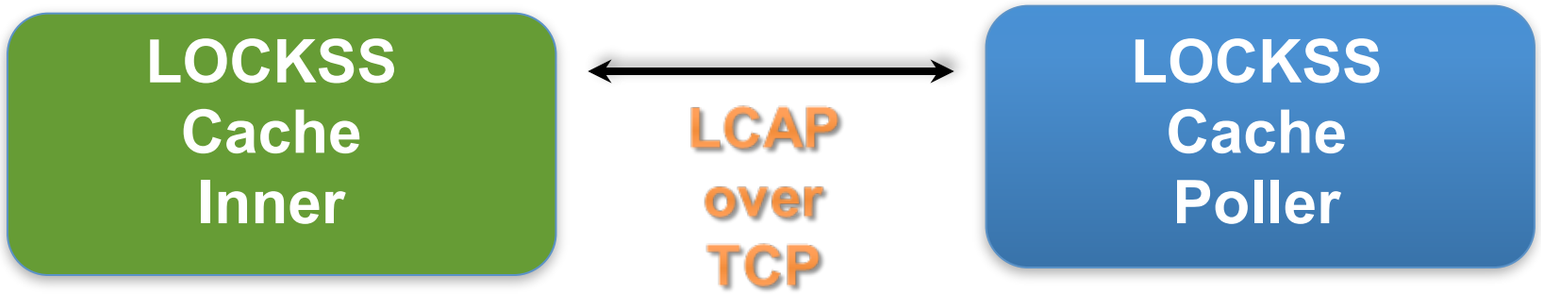


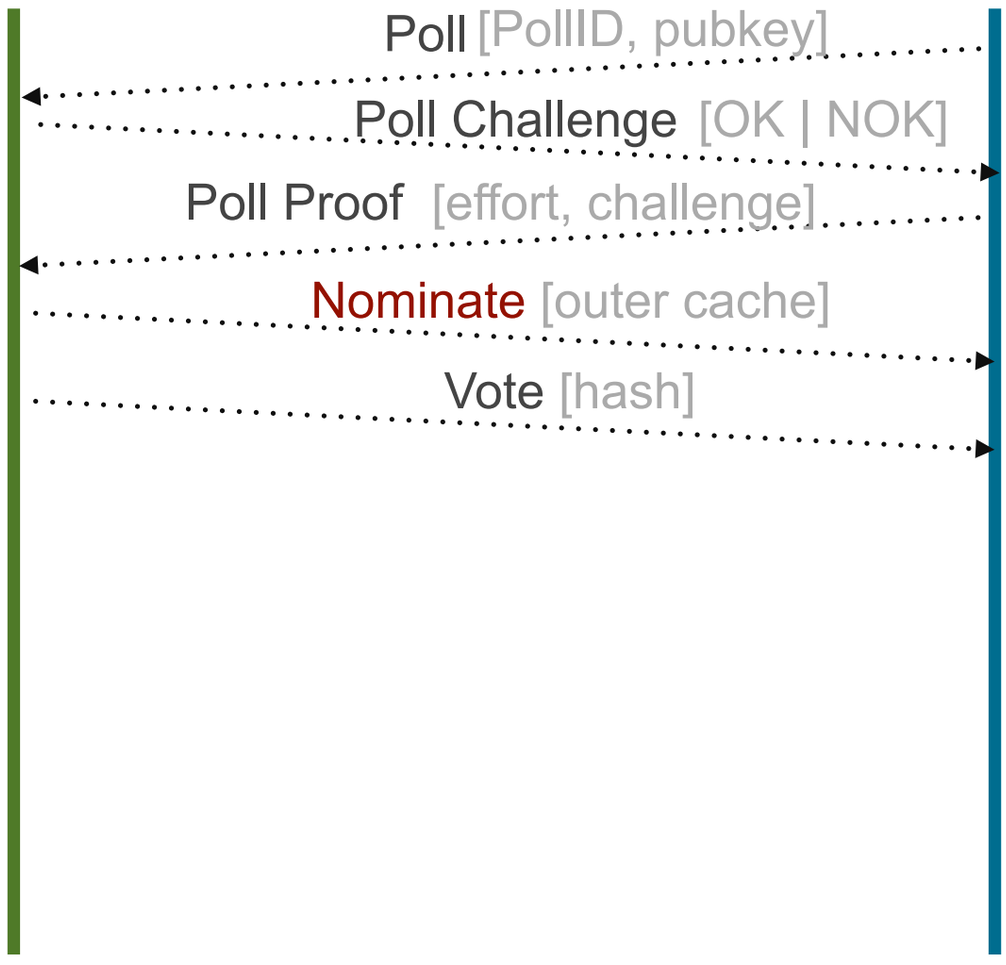
LOCKSS in practice: cache integrity monitoring



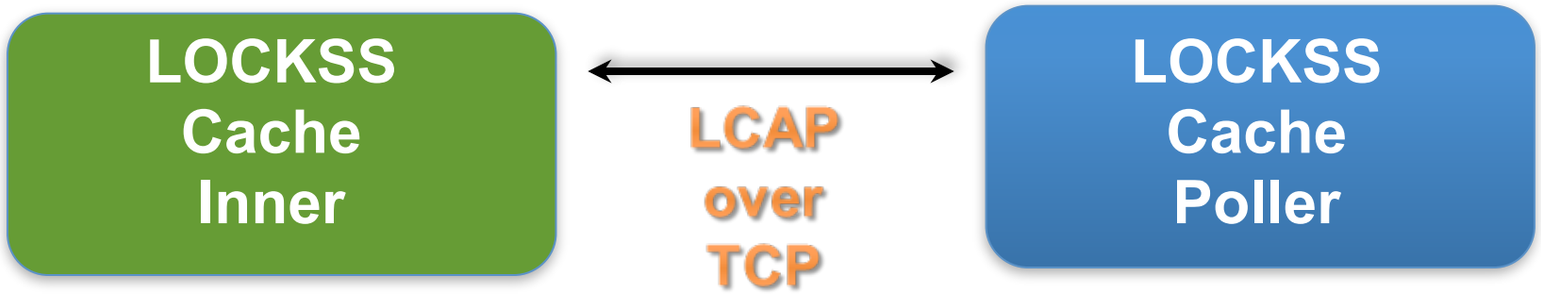


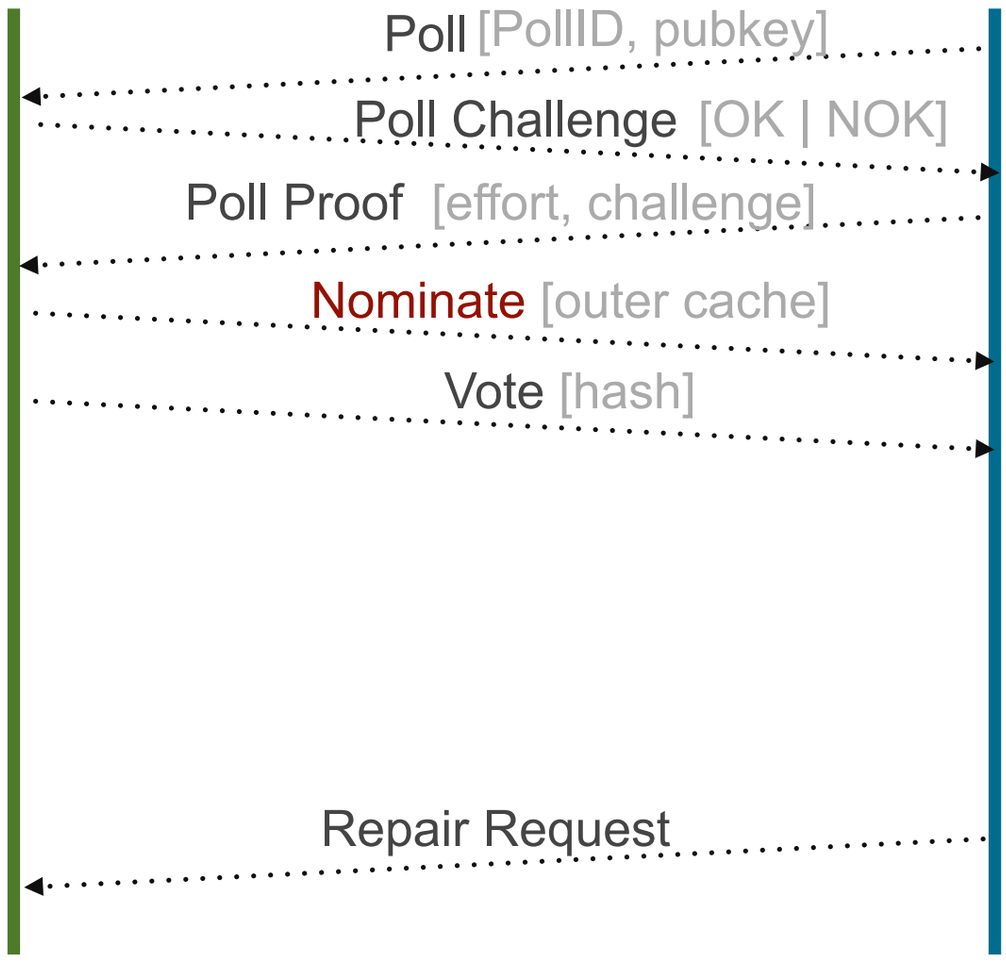
LOCKSS in practice: cache integrity monitoring



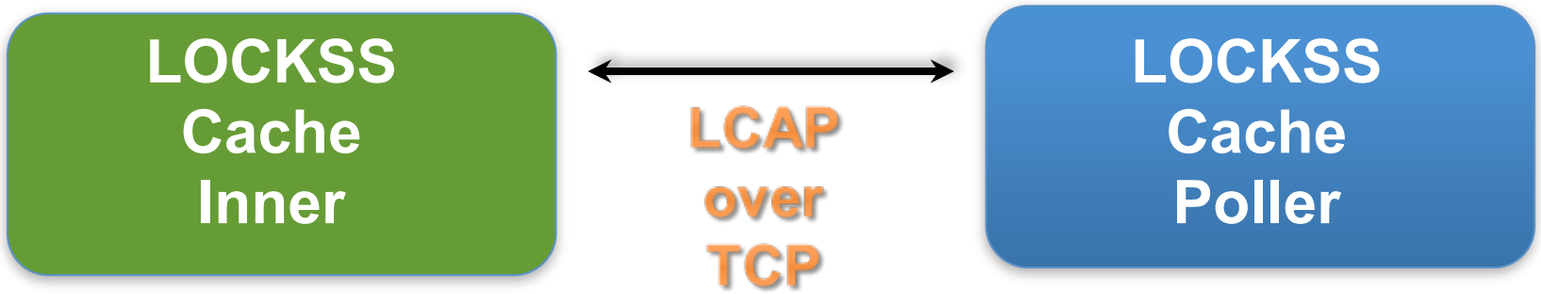


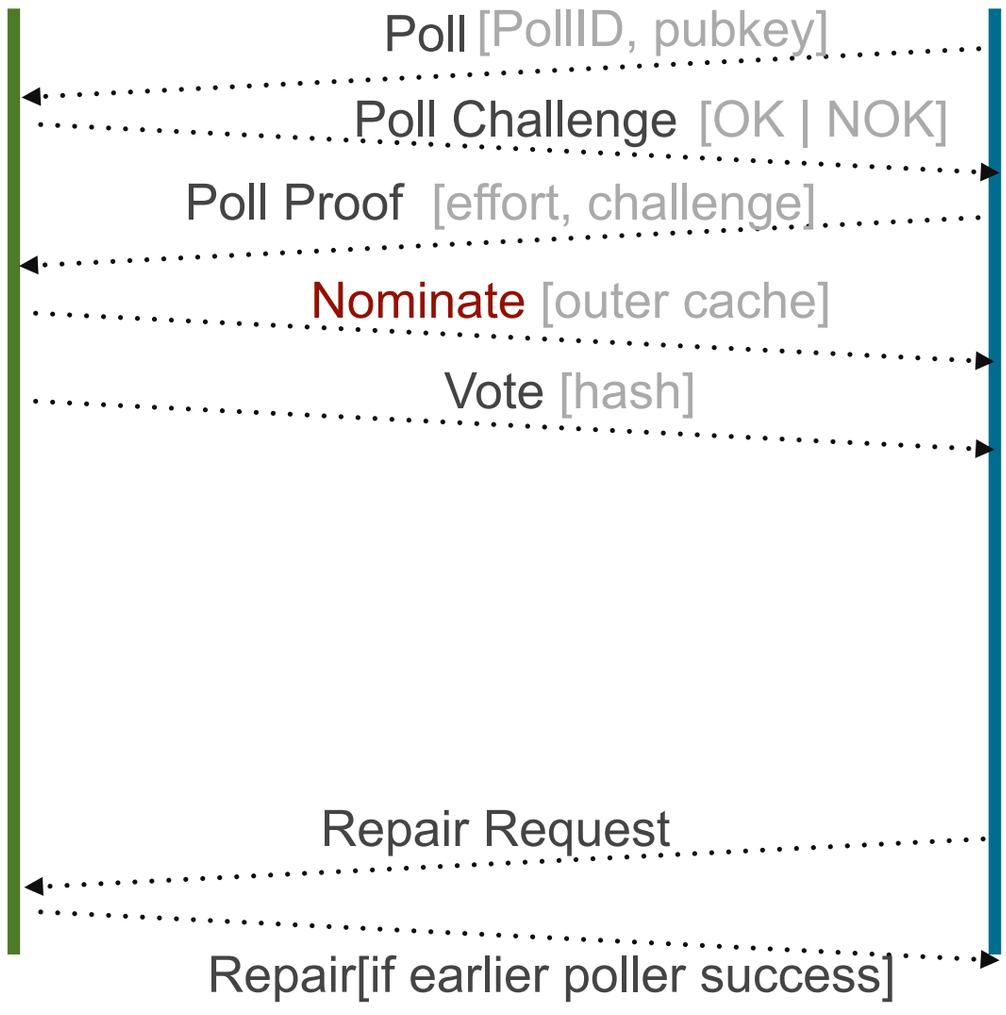
LOCKSS in practice: cache integrity monitoring





LOCKSS in practice: cache integrity monitoring





Agenda

Motivation

LOCKSS main concepts

SAFE LOCKSS network : a Case Study

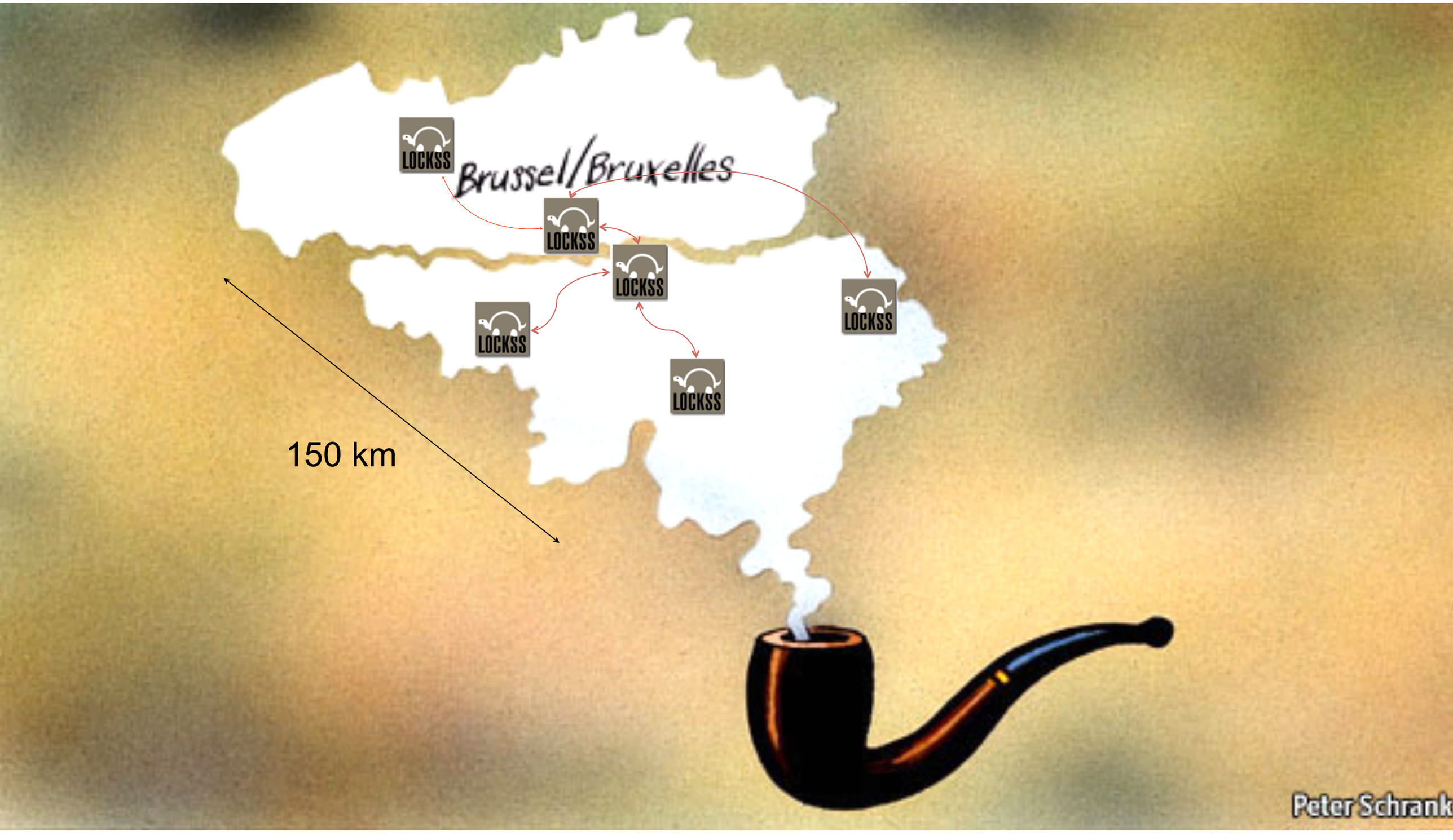
Build your own LOCKSS Network

Discussions

Our project started 6 years ago in Belgium…

Ceci n’est pas un PLN !

Ceci n’est pas un PLN !



Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

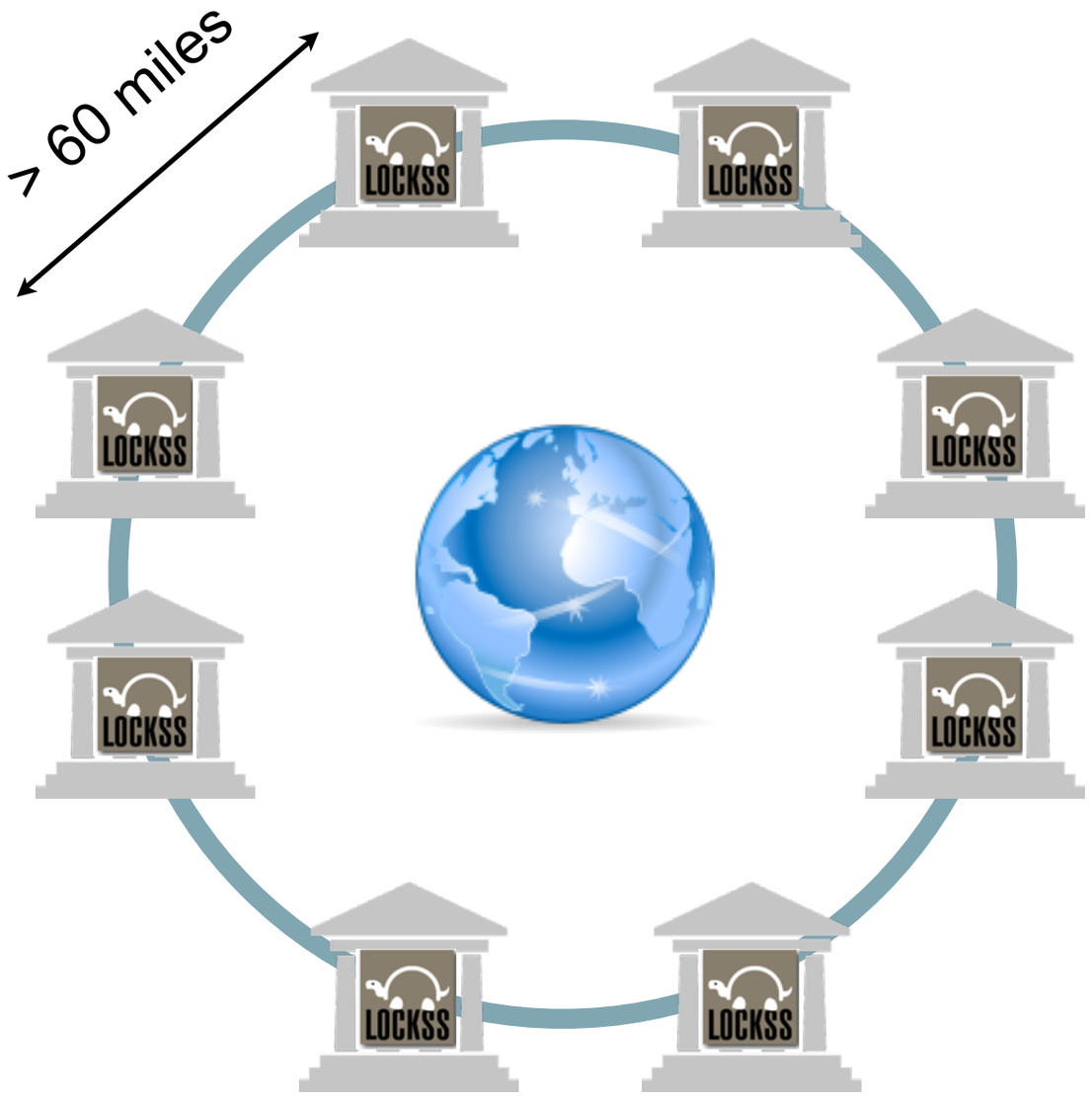
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

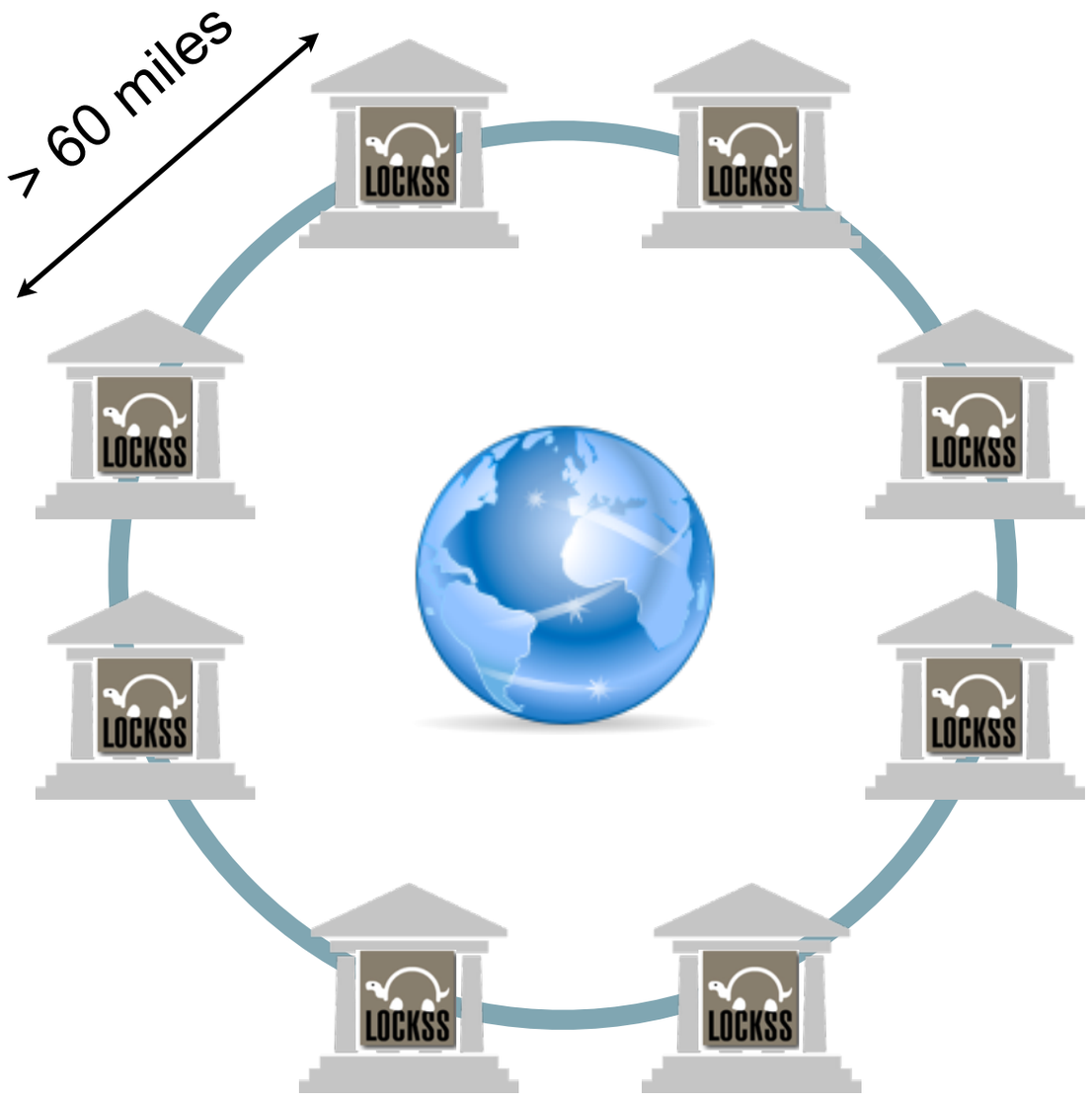
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

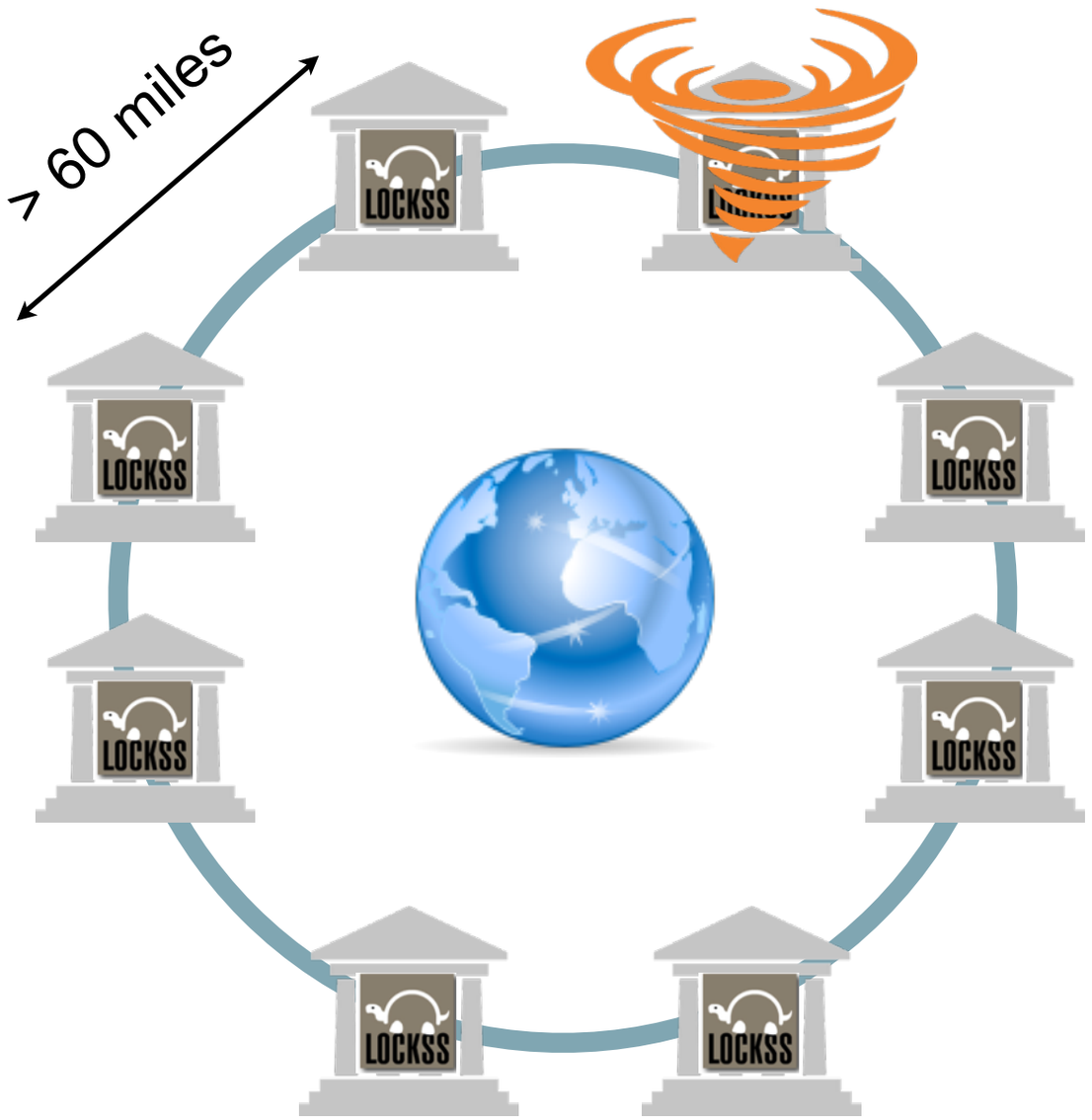
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

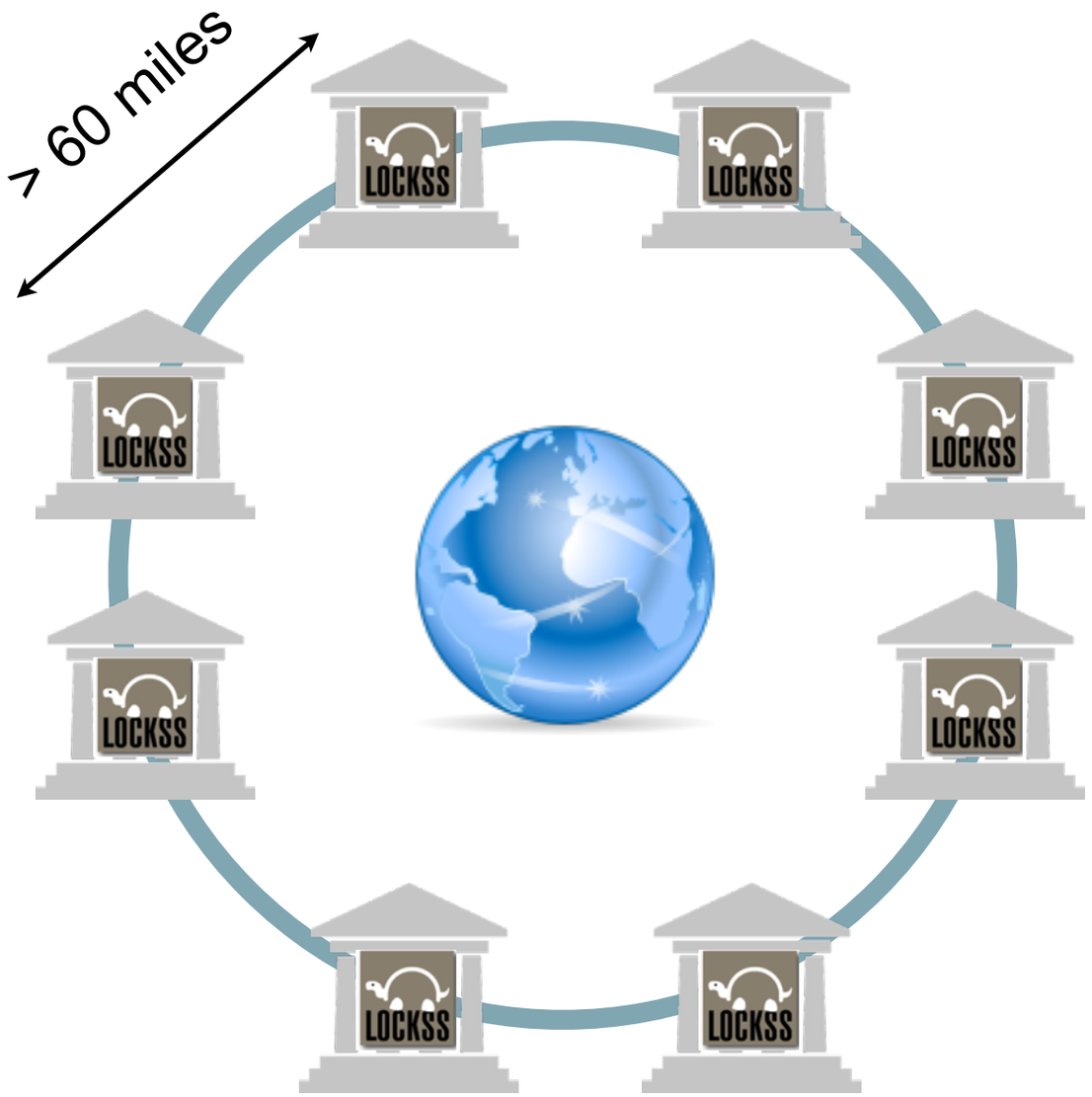
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

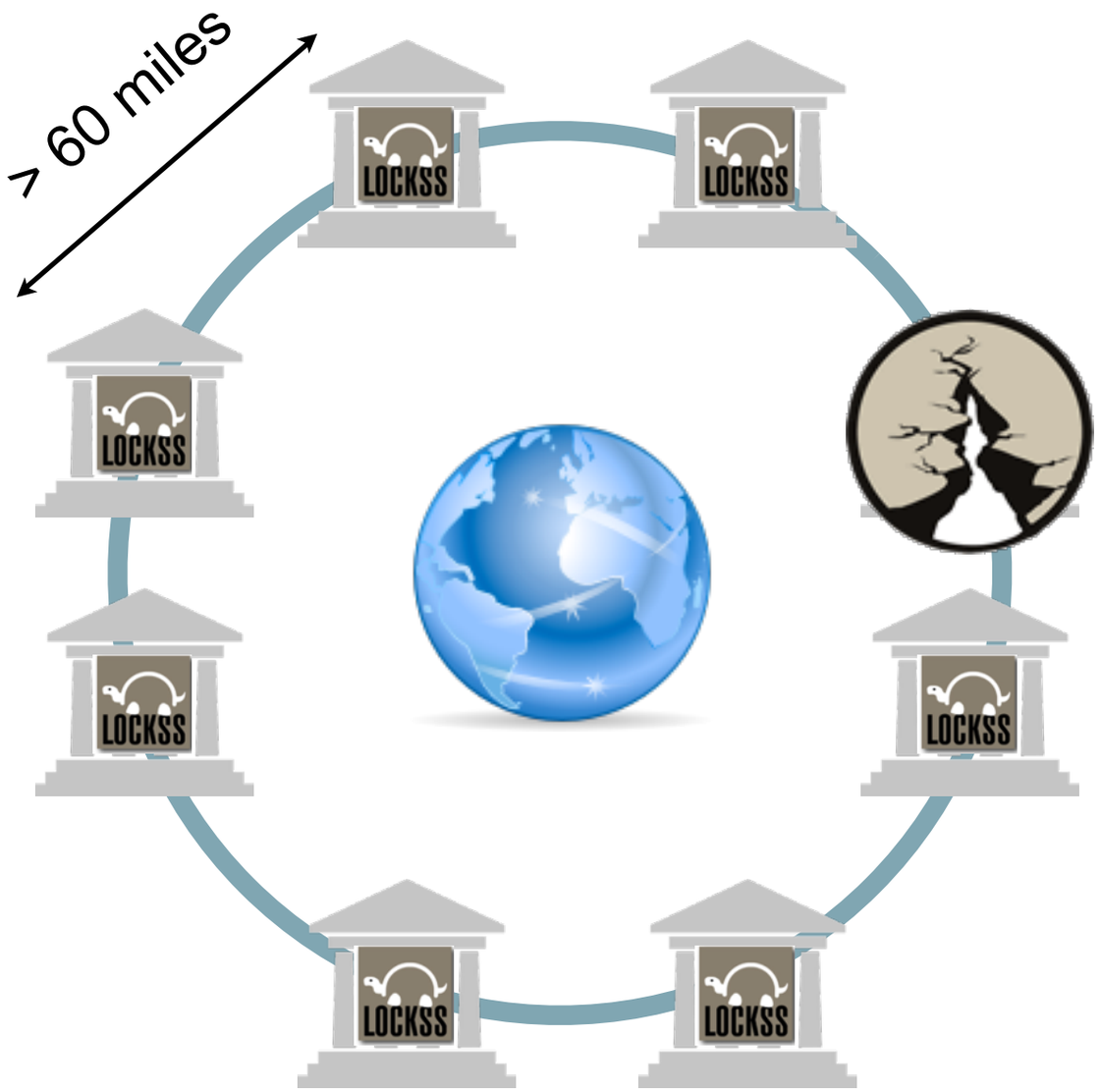
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

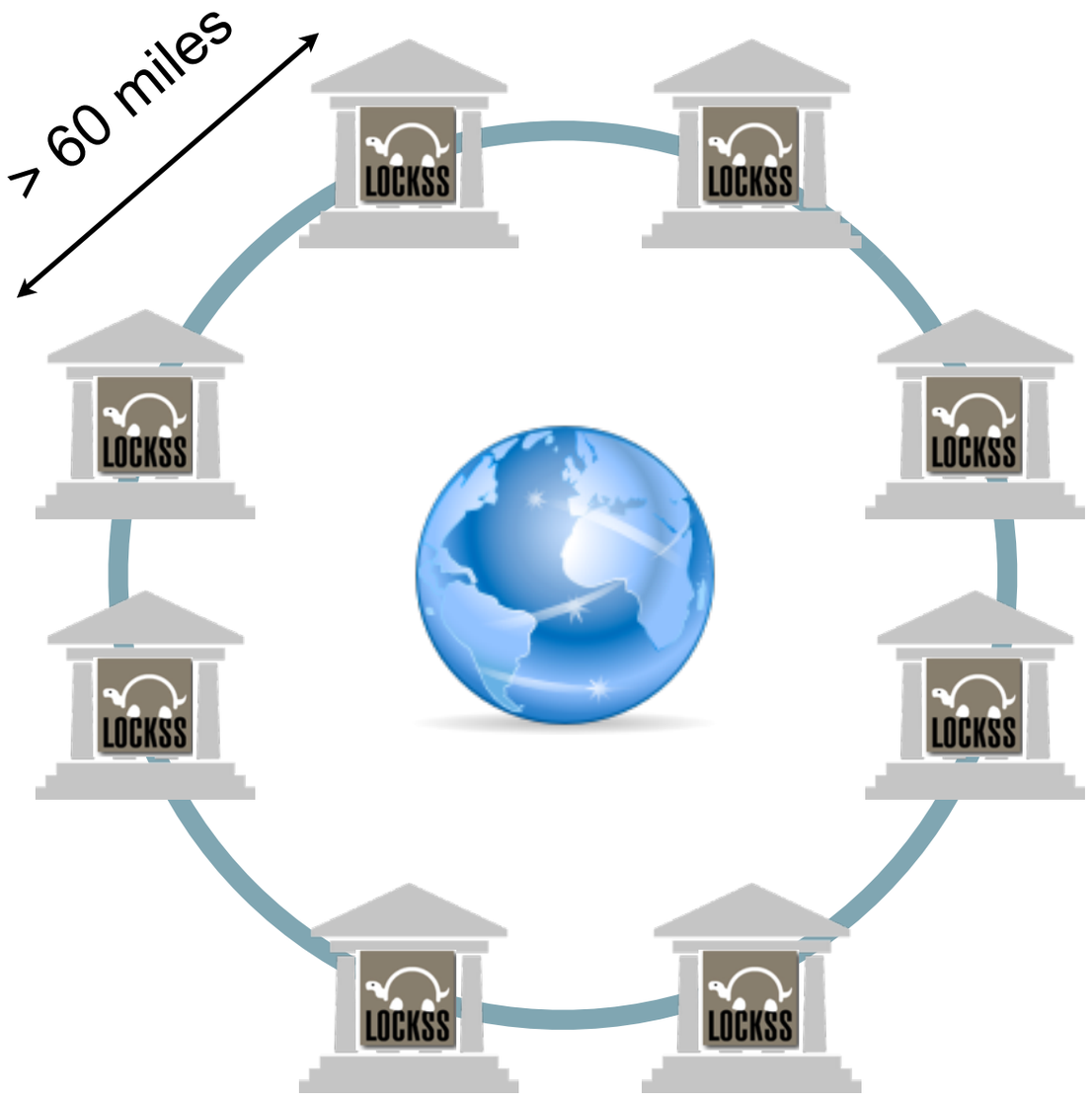
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

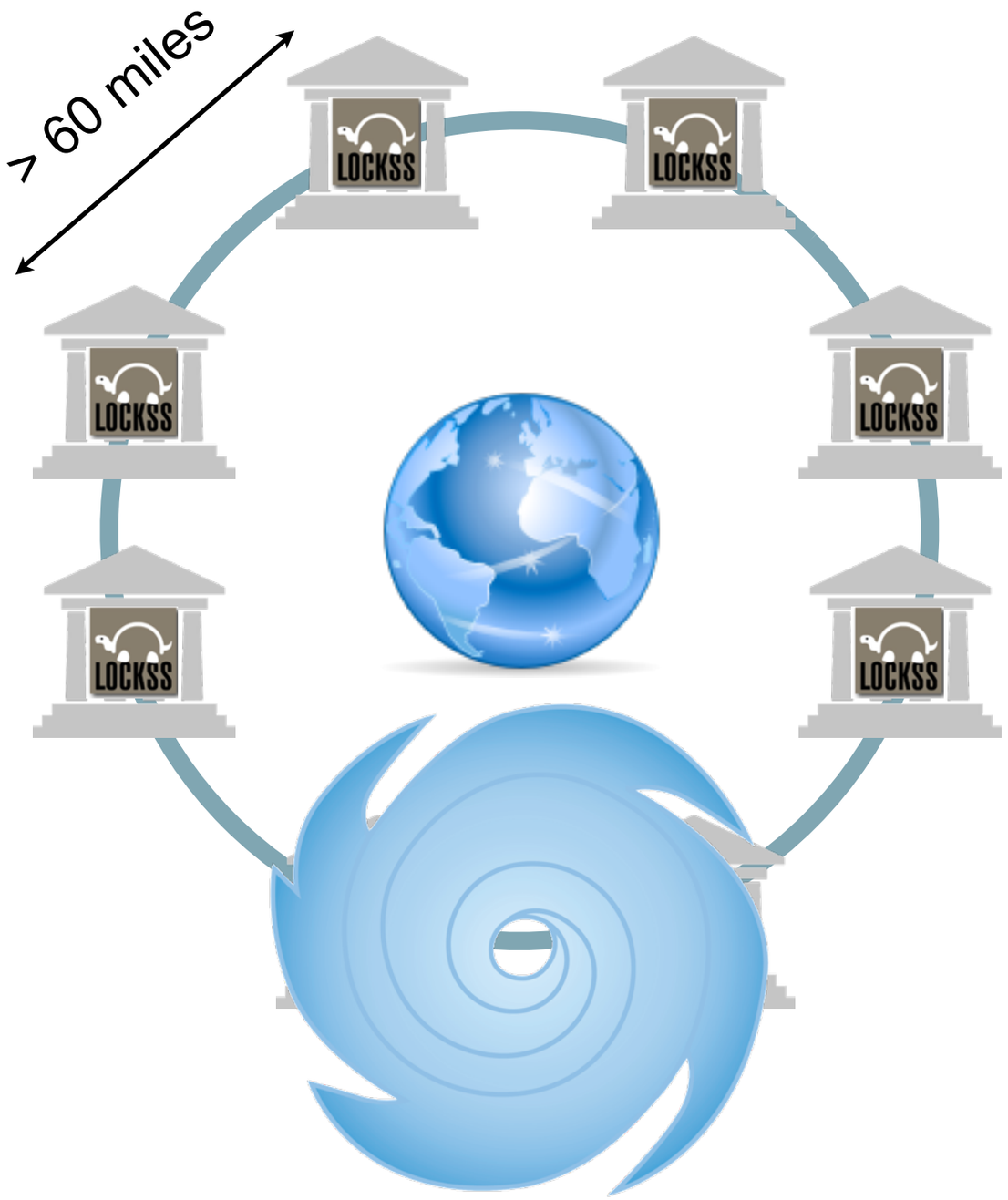
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our preservation solution had to provide proper geo-replication

Cost control

Independent Tech Management

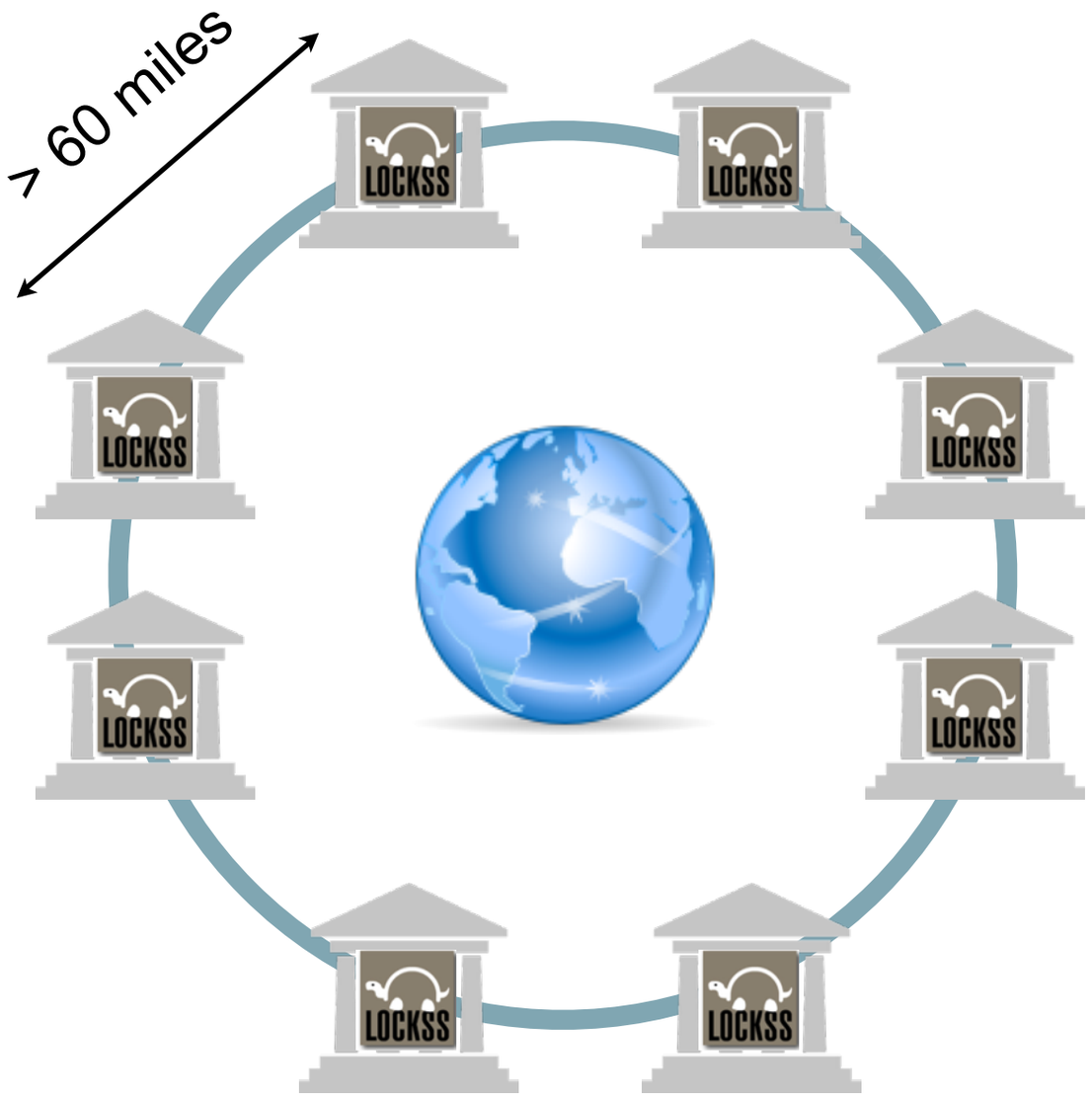
Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring



International collaboration is compulsory for us

Our network had to be based on independent administrations: various institutions under different organizations and governments

Cost control

Independent Tech Management

Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring

Our network had to be based on independent administrations: various institutions under different organizations and governments

Cost control

Independent Tech Management

Geo-replication

Independent Administration

Authentification

Media Migration

Data Monitoring

LOCKSS facilitates international collaboration as it only requires an agreement to share hardware resources and not a global budget.

LEARNING RESOURCES

SAFE PLN

SAFE Archive FEderation

LEARNING RESOURCES

SAFE PLN

SAFE Archive FEderation

international federation

geo-replication in completely independent sites

LEARNING RESOURCES

SAFE PLN

SAFE Archive FEderation

international federation

geo-replication in completely independent sites

light organizational structure

7+ nodes

LEARNING RESOURCES

SAFE PLN

SAFE Archive FEderation

international federation

geo-replication in completely independent sites

light organizational structure

7+ nodes

distributed technical administration

local admin only, no automation

SAFE PLN

SAFE Archive FEderation

international federation

geo-replication in completely independent sites

light organizational structure

7+ nodes

distributed technical administration

local admin only, no automation

each partner monitors the status of his content in the network

global verification that the preservation is performed correctly

LEARNING RESOURCES

SAFE PLN

SAFE Archive FEderation

international federation

geo-replication in completely independent sites

light organizational structure

7+ nodes

distributed technical administration

local admin only, no automation

each partner monitors the status of his content in the network

global verification that the preservation is performed correctly

budgets remain fully independent

economic risk mitigation

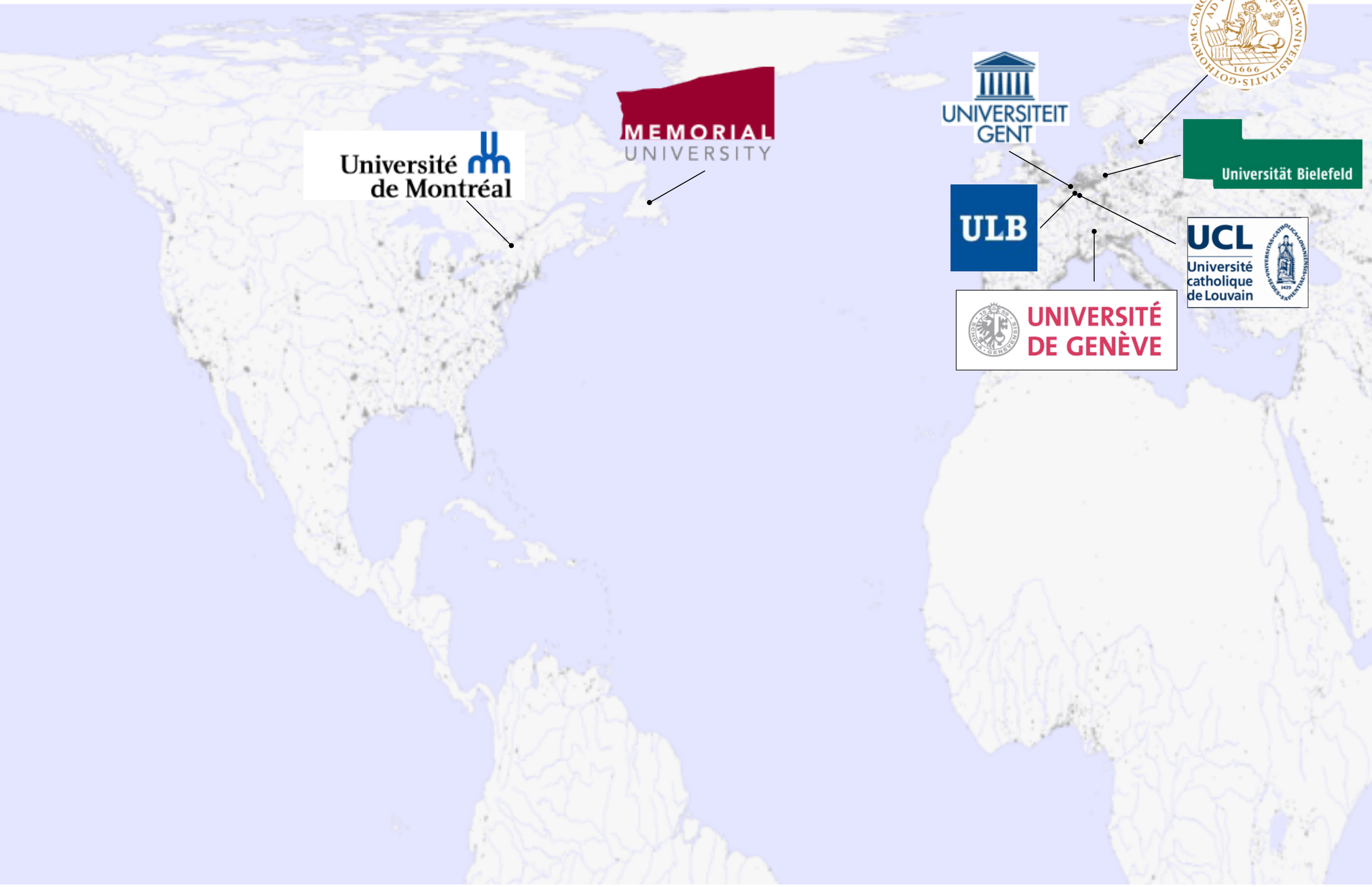
SAFE is based on a simple memorandum of understanding

The parties acknowledge that an **equal amount of resources** will be made available to the execution of the SAFE PLN project. All parties will make a LOCKSS cache node available for the SAFE PLN grid with **a capacity of maximum 4 TB of storage**.

All parties agree that the **maximum raw storage capacity per institution in the SAFE PLN is 400GB**.

The storage capacity requirements will be **adapted every year by mutual agreement**.

SAFE currently has 8 members



SAFE : WHO ?

Université de Montréal

Martin Sévigny

Christian Bélanger

Universität Bielefeld

Friedrich Summann

Christian Pietsch

UNIVERSITEIT GENT

Patrick Hochstenbach

Omar Reygaert

Snorri Briem

Frank Hansen

UCL Université catholique de Louvain

Benoit Erken

Renaud Michotte

MEMORIAL UNIVERSITY

Krista Godfrey

Shawn Wagner

UNIVERSITÉ DE GENÈVE

Jan Melichar

Hugues Cazeaux

Jean-Blaise Claivaz

ULB

Benoit Pauwels

Anthony Leroy

SAFE : WHAT ?

Digitized books
Maps
Pictures

Theses and publications (IR)

Academic content
Research data

SAFE : WHAT ?

Digitized books
Maps
Pictures

Theses and publications (IR)

Academic content
Research data

SAFE : WHAT ?

Université libre de Bruxelles	34 GB
Université catholique de Louvain	8 GB
Bielefeld University	37 GB
Universiteit Gent	261 GB
Université de Genève	25 GB
Lund University	26 GB

Note: Some institutions are still in the process of making their AUs available for production

SAFE : WHICH ?

2012-ETD-ugent: Archival Unit

LEARNING RESOURCES

SAFE : WHICH ?

2012-ETD-ugent: Archival Unit

manifest.html

metadata.xml

SAFE : WHICH ?

2012-ETD-ugent: Archival Unit

manifest.html

metadata.xml

archive-ugent-be-[UUID] : BagIt

- bagit.txt
- bag-info.txt
- manifest-md5.txt
- marc.xml
- mets.xml
- data/
  - file AC.pdf
  - file LS.txt
  - file MA.tar

SAFE : HOW ?

SAFE : HOW ?

The content that is ready to be archived

SAFE : HOW ?

The content that is ready to be archived
is frozen in a staging server.

SAFE : HOW ?

The content that is ready to be archived is frozen in a staging server.

Dissemination repository

SAFE : HOW ?

The content that is ready to be archived is frozen in a staging server.

Dissemination repository

Archive staging server

SAFE : HOW ?

The content that is ready to be archived is frozen in a staging server.

Key LOCKSS aspects for SAFE

“poll/repair” is the key component for us

Key LOCKSS aspects for SAFE

“poll/repair” is the key component for us

share resources, not money

Key LOCKSS aspects for SAFE

“poll/repair” is the key component for us

share resources, not money

distributed technical administration, no automation and centralized administration of boxes

Agenda

Motivation

LOCKSS main concepts

SAFE LOCKSS network : a Case Study

Build your own LOCKSS Network

Discussions

What you need to build your own LOCKSS network

Administrative server

network configuration
[XML]
global parameters

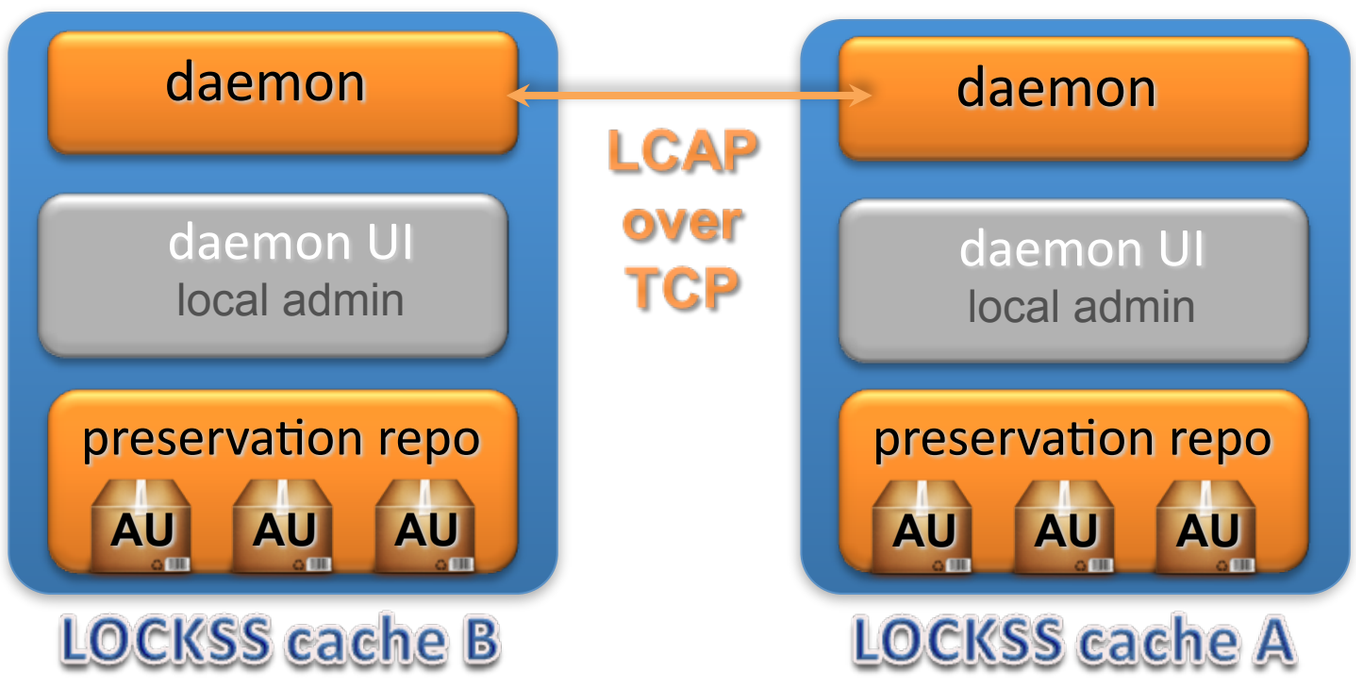
title database
[XML]
Plug-in location
AUs definitions
list of caches IP address

plug-ins repository
CVS for JARs [java | xml]
IR1 certified plug-in
IR2 certified plug-in
IR3 certified plug-in

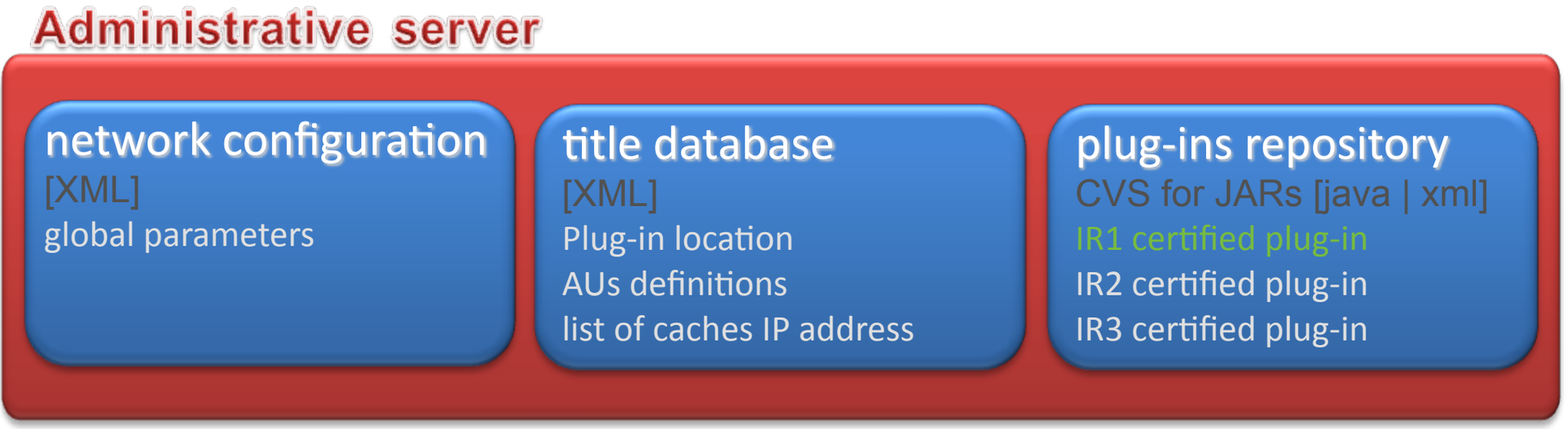
IR1 web server

manifest page
[html]

permissions
[html]



1. Create your administrative server



1. Create your administrative server

Administrative server

network configuration
[XML]
global parameters

title database
[XML]
Plug-in location
AUs definitions
list of caches IP address

plug-ins repository
CVS for JARs [java | xml]
IR1 certified plug-in
IR2 certified plug-in
IR3 certified plug-in

Administration server

./lockss.xml

network configuration &
basic box configuration

./plugins/

how to collect AU?

./titledb/

what to collect in AU?

lockss.xml defines basic box configuration

lockss.xml defines basic box configuration

Global parameters are stored in **lockss.xml**

```
# LOCKSS & LCAP tuning parameters
org.lockss.log.default.level=debug

#lockss config stuff
org.lockss.platform.diskSpacePaths=./
org.lockss.config.reloadInterval = 1m
org.lockss.ui.allowPluginConfig=true

# comm settings - disable V1, V3 only.
org.lockss.ui.start=yes
org.lockss.proxy.start=no
org.lockss.comm.enabled=false

# V3
org.lockss.poll.v3.enableV3Poller=false
org.lockss.poll.defaultPollProtocol=3
org.lockss.scomm.enabled=true
org.lockss.scomm.maxMessageSize=33554430
org.lockss.poll.v3.pollStarterInitialDelay=2m
org.lockss.poll.v3.pollStarterInterval=2m
org.lockss.poll.v3.quorum=2
org.lockss.poll.v3.minNominationSize=1
org.lockss.poll.v3.maxNominationSize=1
org.lockss.id.initialV3PeerList=\
   TCP:[164.15.4.112]:9721;\
   TCP:[164.15.4.113]:9721;

org.lockss.metrics.slowest.hashrate = 250
org.lockss.state.recall.delay=5m

org.lockss.crawler.startCrawlsInterval=5m
```

```
# prod group modes
org.lockss.baseau.defaultFetchRateLimiterSource = plugin
org.lockss.crawler.maxRepairRate = 1000/1d
org.lockss.poll.v3.deleteExtraFiles = true
org.lockss.poll.v3.voteBlockThreshold = 0
org.lockss.repository.globalNodeCache.enabled = true
org.lockss.repository.globalNodeCache.size = 500
org.lockss.state.globalNodeCache.enabled = true
org.lockss.state.globalNodeCache.size = 500
org.lockss.scheduler.maxBackgroundLoad = 50
org.lockss.state.pollHistory.trimRewrite = true
org.lockss.treewalk.v1.mode = off

# Standard debugging settings
org.lockss.monitor.inputStreams=true
org.lockss.thread.hungThreadDump = true

org.lockss.crawlStatus.keepUrls=all

# UI username-passwd = lockss-u/lockss-p
org.lockss.platform.ui.username=lockss-u
org.lockss.platform.ui.password=SHA1:---

# So that we don't hash on no-longer-relevant URLs
org.lockss.blockHasher.ignoreFilesOutsideCrawlSpec=true

org.lockss.baseau.minFetchDelay=4000
org.lockss.plugin.restartAusWithNewPlugin=true
```

lockss.xml defines basic box configuration

Global parameters are stored in **lockss.xml**

```
# LOCKSS & LCAP tuning parameters
org.lockss.log.default.level=debug

#lockss config stuff
org.lockss.platform.diskSpacePaths=./
org.lockss.config.reloadInterval = 1m
org.lockss.ui.allowPluginConfig=true

# comm settings - disable V1, V3 only.
org.lockss.ui.start=yes
org.lockss.proxy.start=no
org.lockss.comm.enabled=false

# V3
org.lockss.poll.v3.enableV3Poller=false
org.lockss.poll.defaultPollProtocol=3
org.lockss.scomm.enabled=true
org.lockss.scomm.maxMessageSize=33554430
org.lockss.poll.v3.pollStarterInitialDelay=2m
org.lockss.poll.v3.pollStarterInterval=2m
org.lockss.poll.v3.quorum=2
org.lockss.poll.v3.minNominationSize=1
org.lockss.poll.v3.maxNominationSize=1
org.lockss.id.initialV3PeerList=\
    TCP:[164.15.4.112]:9721;\
    TCP:[164.15.4.113]:9721;

org.lockss.metrics.slowest.hashrate = 250
org.lockss.state.recall.delay=5m

org.lockss.crawler.startCrawlsInterval=5m
```

```
# prod group modes
org.lockss.baseau.defaultFetchRateLimiterSource = plugin
org.lockss.crawler.maxRepairRate = 1000/1d
org.lockss.poll.v3.deleteExtraFiles = true
org.lockss.poll.v3.voteBlockThreshold = 0
org.lockss.repository.globalNodeCache.enabled = true
org.lockss.repository.globalNodeCache.size = 500
org.lockss.state.globalNodeCache.enabled = true
org.lockss.state.globalNodeCache.size = 500
org.lockss.scheduler.maxBackgroundLoad = 50
org.lockss.state.pollHistory.trimRewrite = true
org.lockss.treewalk.v1.mode = off

# Standard debugging settings
org.lockss.monitor.inputStreams=true
org.lockss.thread.hungThreadDump = true

org.lockss.crawlStatus.keepUrls=all

# UI username-passwd = lockss-u/lockss-p
org.lockss.platform.ui.username=lockss-u
org.lockss.platform.ui.password=SHA1:---

# So that we don't hash on no-longer-relevant URLs
org.lockss.blockHasher.ignoreFilesOutsideCrawlSpec=true

org.lockss.baseau.minFetchDelay=4000
org.lockss.plugin.restartAusWithNewPlugin=true
```

Daemon-specific parameters are set in **/etc/lockss/config.dat**

```
org.lockss.proxy.start=yes
org.lockss.ui.port=8081

#test settings
org.lockss.test.doTreewalk=false
org.lockss.test.treewalk.auId=org|lockss|plugin|simulated|
SimulatedPlugin&root~localA

org.lockss.test.doPoll=false
org.lockss.test.polltype=2
org.lockss.test.pollspec.auId=org|lockss|plugin|simulated|
SimulatedPlugin&root~localA
org.lockss.test.pollspec.url=lockssau:
#org.lockss.test.pollspec.lwrBound=
#org.lockss.test.pollspec.uprBound=

org.lockss.proxy.icp.enabled=true
org.lockss.proxy.icp.port=3131

org.lockss.platform.v3.identity=TCP:[127.0.0.1]:9721
```

lockss.xml defines basic box configuration

Global parameters are stored in **lockss.xml**

```
# LOCKSS & LCAP tuning parameters
org.lockss.log.default.level=debug

#lockss config stuff
org.lockss.platform.diskSpacePaths=./
org.lockss.config.reloadInterval = 1m
org.lockss.ui.allowPluginConfig=true

# comm settings - disable V1, V3 only.
org.lockss.ui.start=yes
org.lockss.proxy.start=no
org.lockss.comm.enabled=false

# V3
org.lockss.poll.v3.enableV3Poller=false
org.lockss.poll.defaultPollProtocol=3
org.lockss.scomm.enabled=true
org.lockss.scomm.maxMessageSize=33554430
org.lockss.poll.v3.pollStarterInitialDelay=2m
org.lockss.poll.v3.pollStarterInterval=2m
org.lockss.poll.v3.quorum=2
org.lockss.poll.v3.minNominationSize=1
org.lockss.poll.v3.maxNominationSize=1
org.lockss.id.initialV3PeerList=\
   TCP:[164.15.4.112]:9721;\
   TCP:[164.15.4.113]:9721;

org.lockss.metrics.slowest.hashrate = 250
org.lockss.state.recall.delay=5m

org.lockss.crawler.startCrawlsInterval=5m
```

```
# prod group modes
org.lockss.baseau.defaultFetchRateLimiterSource = plugin
org.lockss.crawler.maxRepairRate = 1000/1d
org.lockss.poll.v3.deleteExtraFiles = true
org.lockss.poll.v3.voteBlockThreshold = 0
org.lockss.repository.globalNodeCache.enabled = true
org.lockss.repository.globalNodeCache.size = 500
org.lockss.state.globalNodeCache.enabled = true
org.lockss.state.globalNodeCache.size = 500
org.lockss.scheduler.maxBackgroundLoad = 50
org.lockss.state.pollHistory.trimRewrite = true
org.lockss.treewalk.v1.mode = off

# Standard debugging settings
org.lockss.monitor.inputStreams=true
org.lockss.thread.hungThreadDump = true

org.lockss.crawlStatus.keepUrls=all

# UI username-passwd = lockss-u/lockss-p
org.lockss.platform.ui.username=lockss-u
org.lockss.platform.ui.password=SHA1:---

# So that we don't hash on no-longer-relevant URLs
org.lockss.blockHasher.ignoreFilesOutsideCrawlSpec=true

org.lockss.baseau.minFetchDelay=4000
org.lockss.plugin.restartAusWithNewPlugin=true
```

Daemon-specific parameters are set in **/etc/lockss/config.dat**

```
org.lockss.proxy.start=yes
org.lockss.ui.port=8081

#test settings
org.lockss.test.doTreewalk=false
org.lockss.test.treewalk.auId=org|lockss|plugin|simulated|
SimulatedPlugin&root~localA

org.lockss.test.doPoll=false
org.lockss.test.polltype=2
org.lockss.test.pollspec.auId=org|lockss|plugin|simulated|
SimulatedPlugin&root~localA
org.lockss.test.pollspec.url=lockssau:
#org.lockss.test.pollspec.lwrBound=
#org.lockss.test.pollspec.uprBound=

org.lockss.proxy.icp.enabled=true
org.lockss.proxy.icp.port=3131

org.lockss.platform.v3.identity=TCP:[127.0.0.1]:9721
```

Overriding parameters

Other parameters can be set at the level of the Archival Unit or at the plug-in level

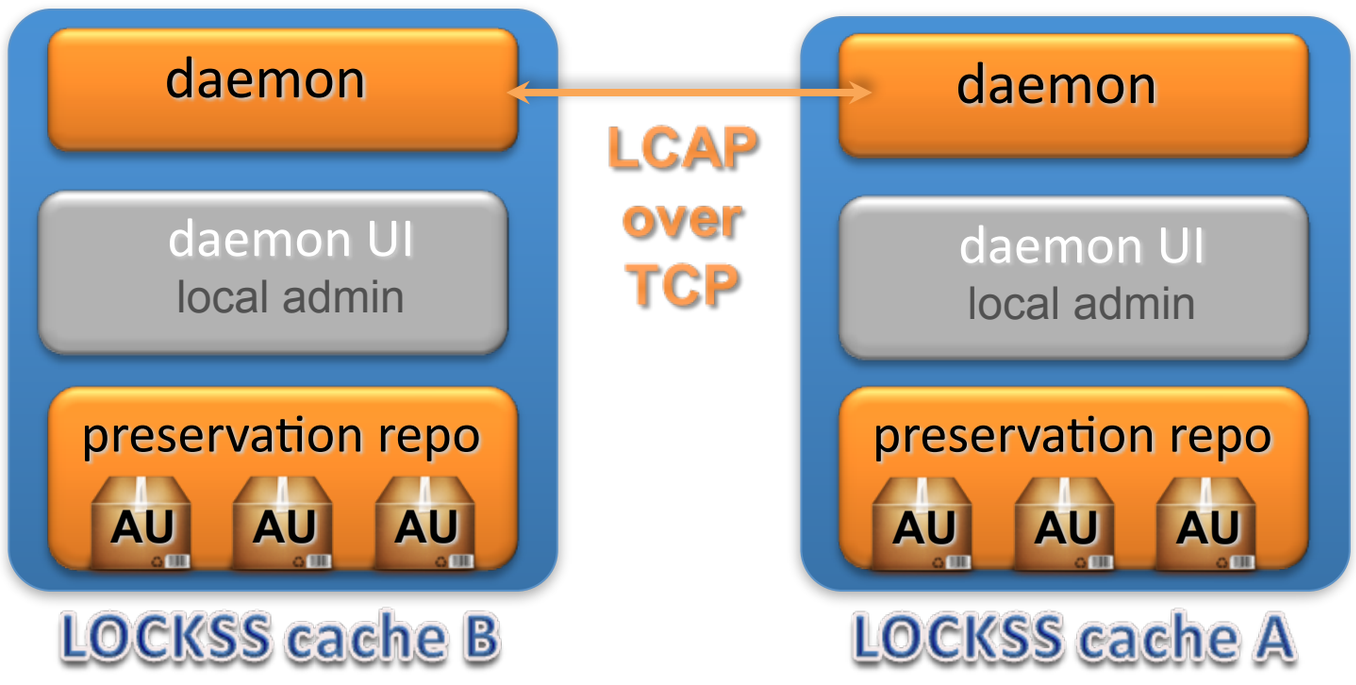
2. Create your LOCKSS box

Administrative server

network configuration
[XML]
global parameters

title database
[XML]
Plug-in location
AUs definitions
list of caches IP address

plug-ins repository
CVS for JARs [java | xml]
IR1 certified plug-in
IR2 certified plug-in
IR3 certified plug-in



What hardware do you need?

LOCKSS cache technical specifications

Principle

7 copies on LOCKSS commodity hardware is:

- much less expensive than *a single* copy on EMC storage
- most importantly much more reliable

Dell PowerEdge R515 - 12x 3TB HDD

Better for preservation if caches are heterogeneous (various hardware configurations)

Building your own LOCKSS box: software requirements

Prerequisite: CentOS 6 basic server install

Open TCP ports 8080 (proxy), 8081 (webadmin), 9729 (V3 poll) and 22 (ssh):

```
+ iptables -A INPUT -m state --state NEW -m tcp -p tcp --dport 22 -j ACCEPT
+ iptables -A INPUT -m state --state NEW -m tcp -p tcp --dport 8080 -j ACCEPT
+ iptables -A INPUT -m state --state NEW -m tcp -p tcp --dport 8081 -j ACCEPT
+ iptables -A INPUT -m state --state NEW -m tcp -p tcp --dport 9729 -j ACCEPT
```

Getting and installing the LOCKSS daemon in five easy steps

1. Log in as root to your lockss cache

```
$ ssh root@[your_lockss_box_ip_address]
```

2. Create a new file, lockss.repo under /etc/yum.repos.d containing the following lines:

```
[lockss]
name = LOCKSS Daemon Repository
baseurl=http://www.lockss.org/repo/
gpgcheck=1
enabled=1
```

3. Install the LOCKSS RPM GPG key:

```
$ rpm --import http://www.lockss.org/LOCKSS-GPG-RPM-KEY
```

4. Installed the LOCKSS daemon with:

```
$ yum install lockss-daemon
```

5. Create local cache folder

```
$ mkdir /cache
```

Configuring your LOCKSS daemon

1. Run the LOCKSS daemon configuration script

   > /etc/lockss/hostconfig

2. Answer local configuration questions as follows

   Fully qualified hostname (FQDN) of this machine: [lockssbox1.ulb.ac.be]
   IP address of this machine: [164.15.4.112]
   Is this machine behind NAT?: [N]
   Initial subnet for admin UI access: [164.15.4.0/24]
   LCAP V3 protocol port: [9729]
   PROXY port: [8080]
   Admin UI port: [8081]
   Mail relay for this machine: [localhost] smtp.ulb.ac.be
   Does mail relay smtp.ulb.ac.be need user & password: [N] y
   User for smtp_server: [] lockssadmin@ulb.ac.be
   Password for locksadmin@ulb.ac.be @ smtp.ulb.ac.be: []
   Again: []
   E-mail address for administrator: [] lockssadmin@ulb.ac.be
   Path to java: [/usr/bin/java]
   Java switches: []
   Configuration URL: [http://props.lockss.org:8001/daemon/lockss.xml] http://164.15.1.1/lockss.xml
   Configuration proxy (host:port): [NONE]
   Preservation group(s): [prod] testpln
   Content storage directories: [] /cache
   Temporary storage directory: [/cache/cache/tmp]
   User name for web UI administration: [] lockss
   Password for web UI administration user lockss: []
   Password for web UI administration (again): []
   OK to store this configuration: [Y] y

And... we have lift-off !

1. Start the daemon

   > /etc/init.d/lockss start

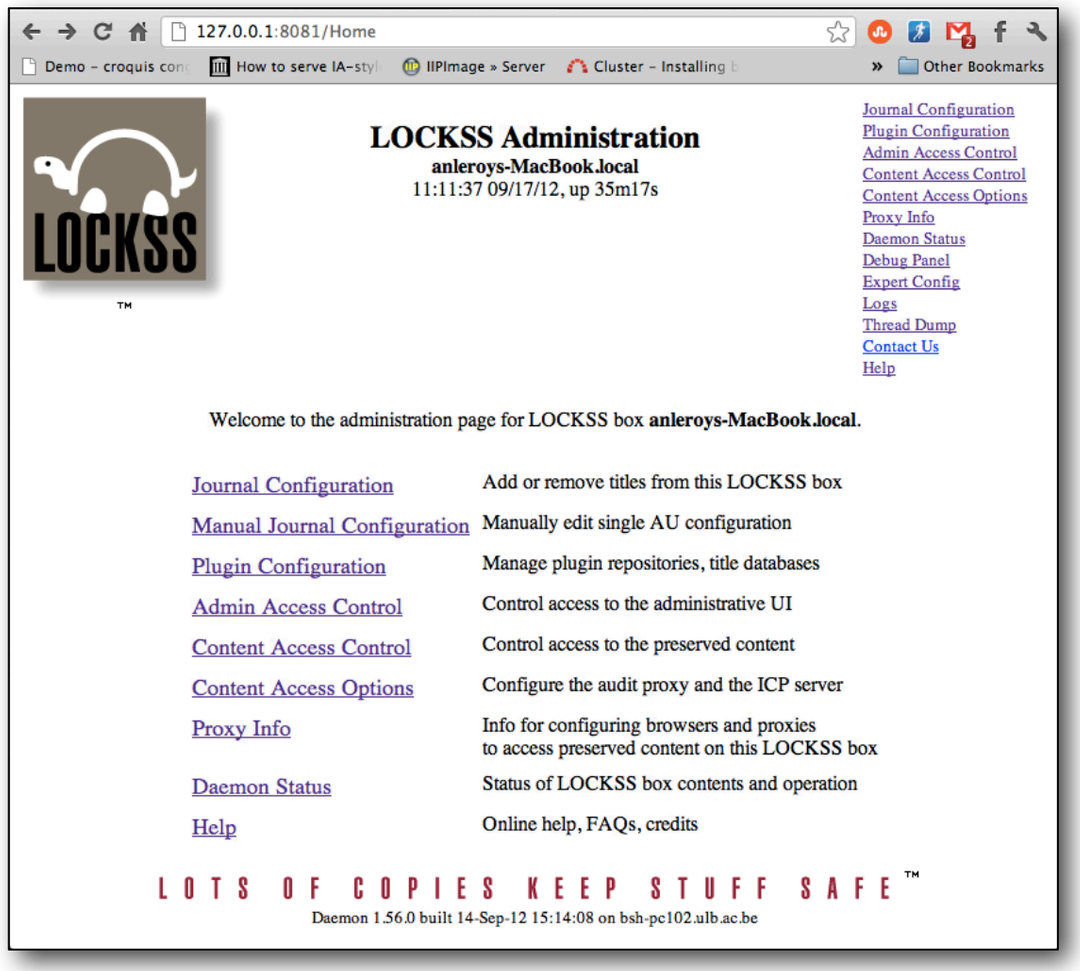
2. Connect to the web interface

   with your favorite browser, go to:
   http://your_ip_address:8081

   user/pwd: previously defined credentials

   no need to chkconfig on, done automatically

   if any error appears, check /var/log/lockss/stdout for info



3. Ask your LOCKSS network admin to update

   - the lockss.xml file with the new cache IP address

3. Create your archive staging server

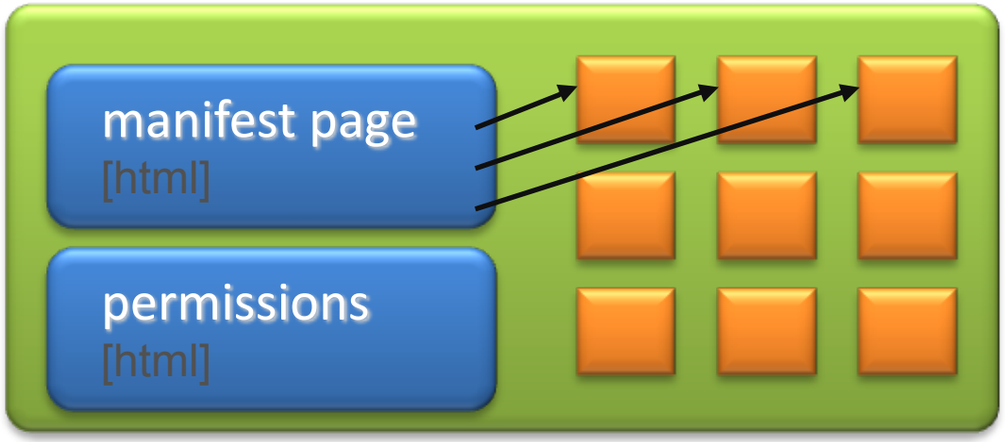
Administrative server

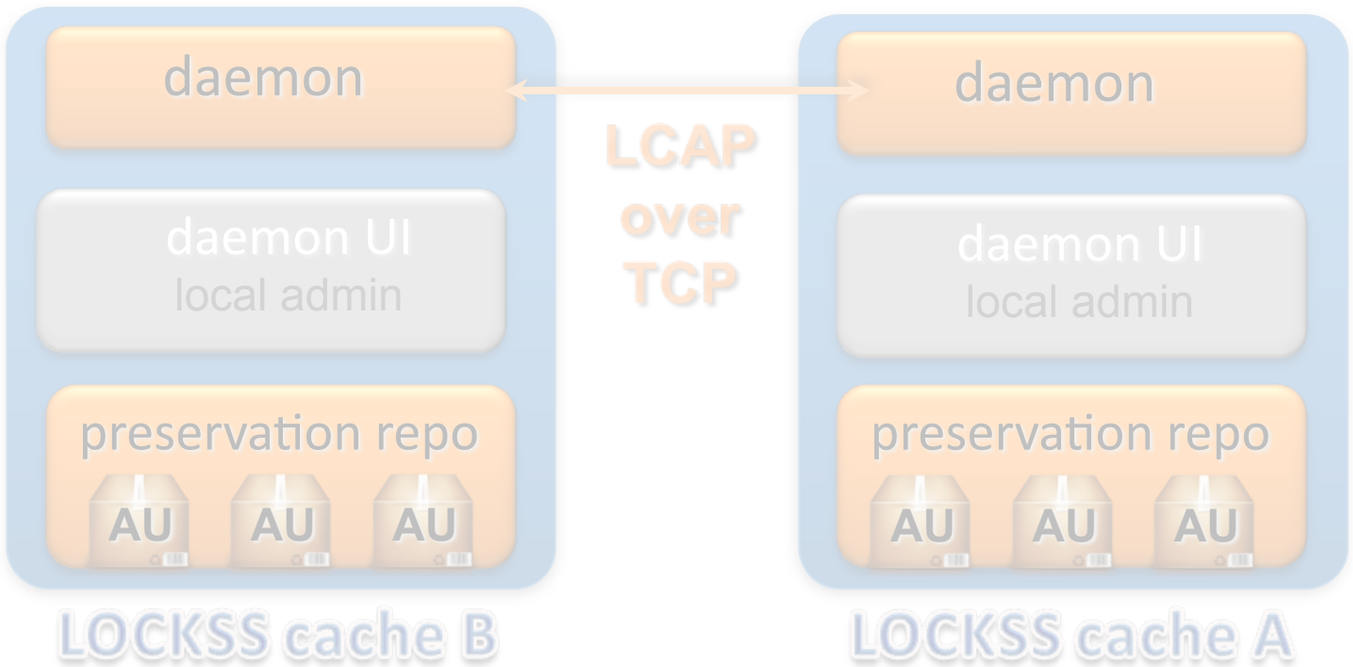
network configuration
[XML]
global parameters

title database
[XML]
Plug-in location
AUs definitions
list of caches IP address

plug-ins repository
CVS for JARs [java | xml]
IR1 certified plug-in
IR2 certified plug-in
IR3 certified plug-in

IR1 web server





To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)

Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

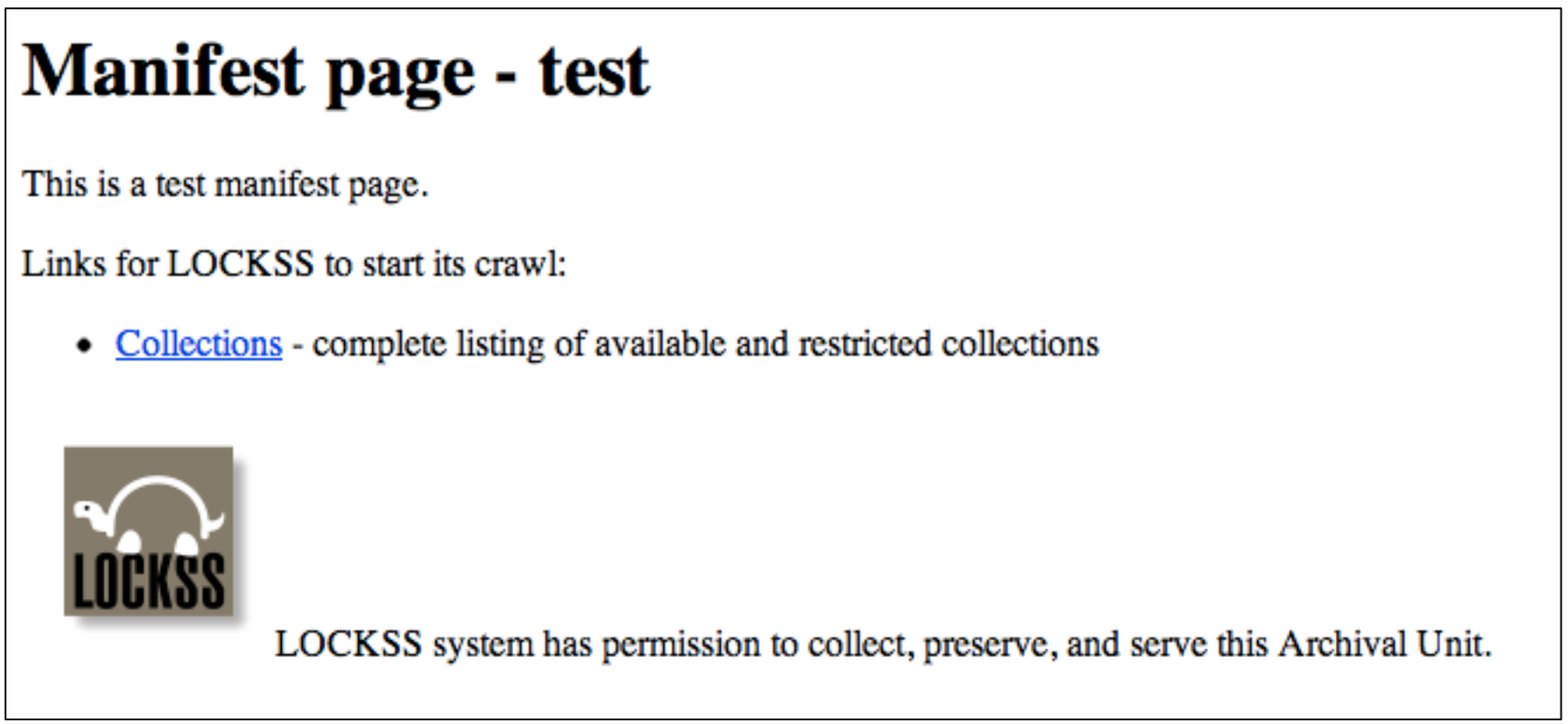
LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)

./lockss/openaccess/openaccess.html (manifest page for the open-access collection)



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

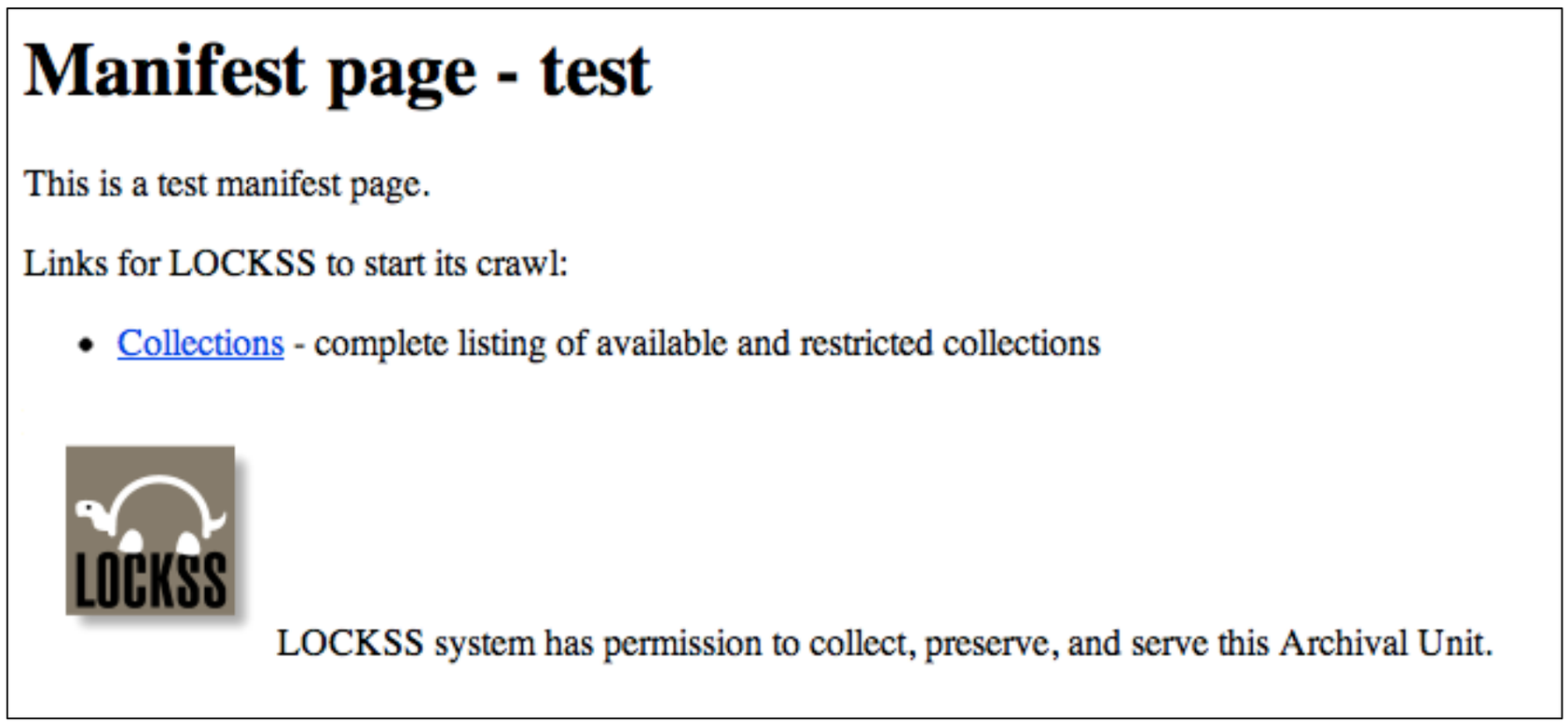
- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

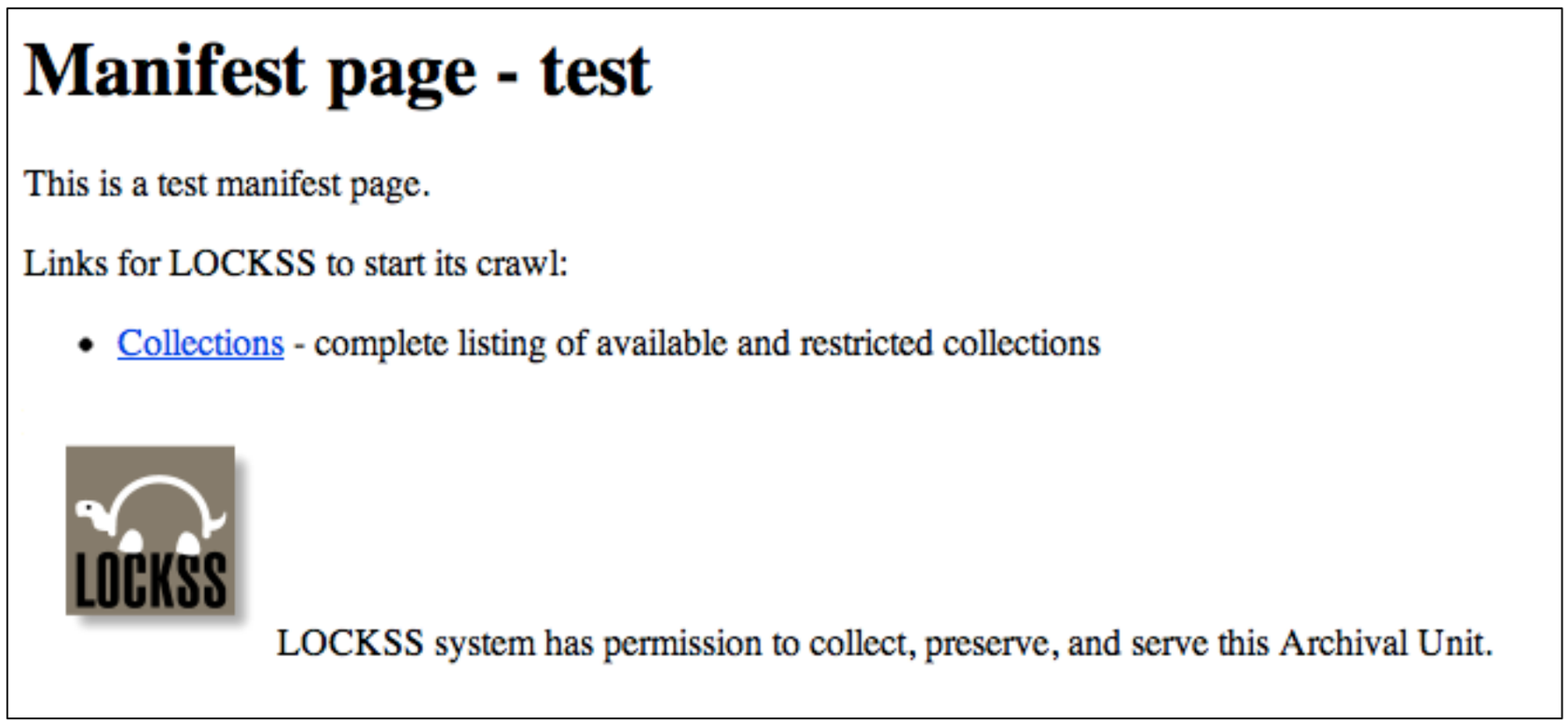
- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

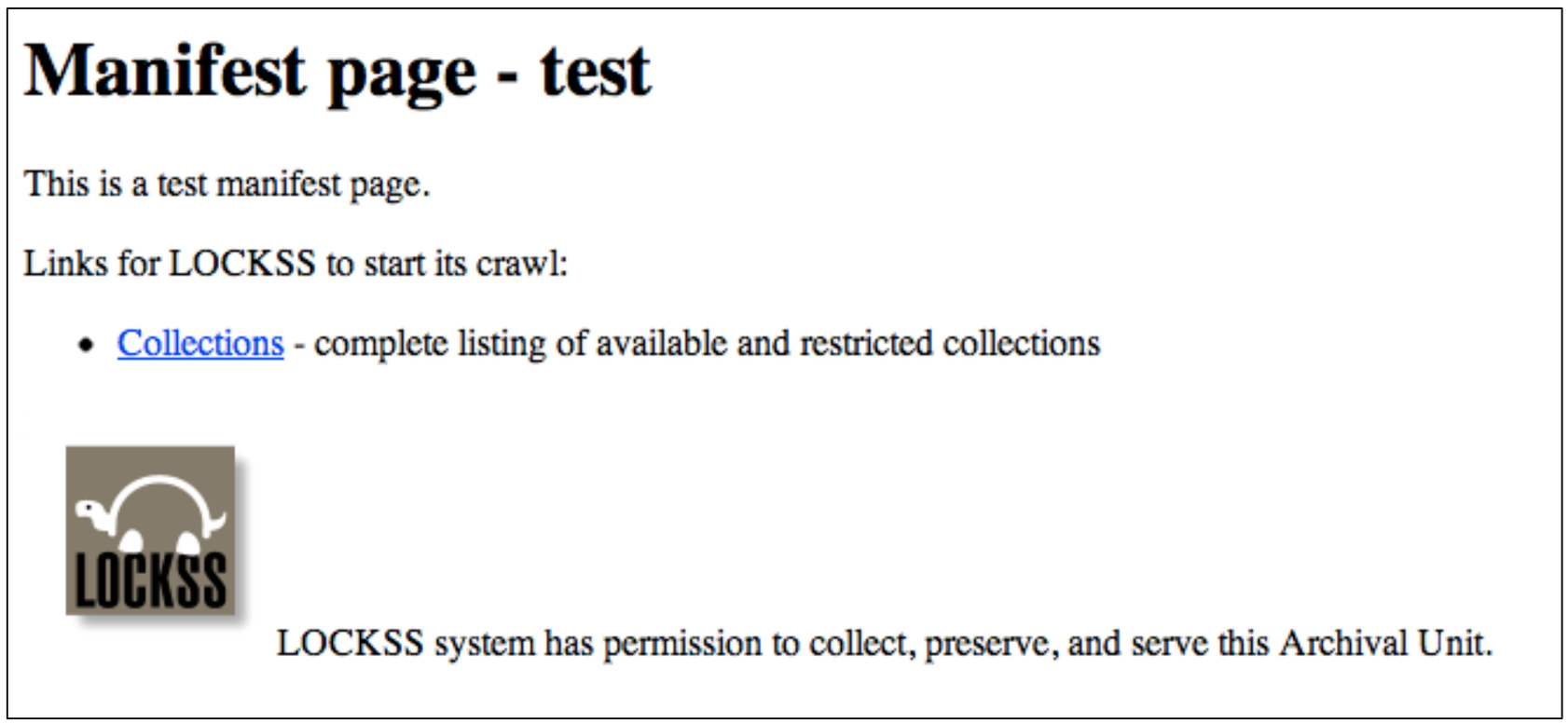
- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

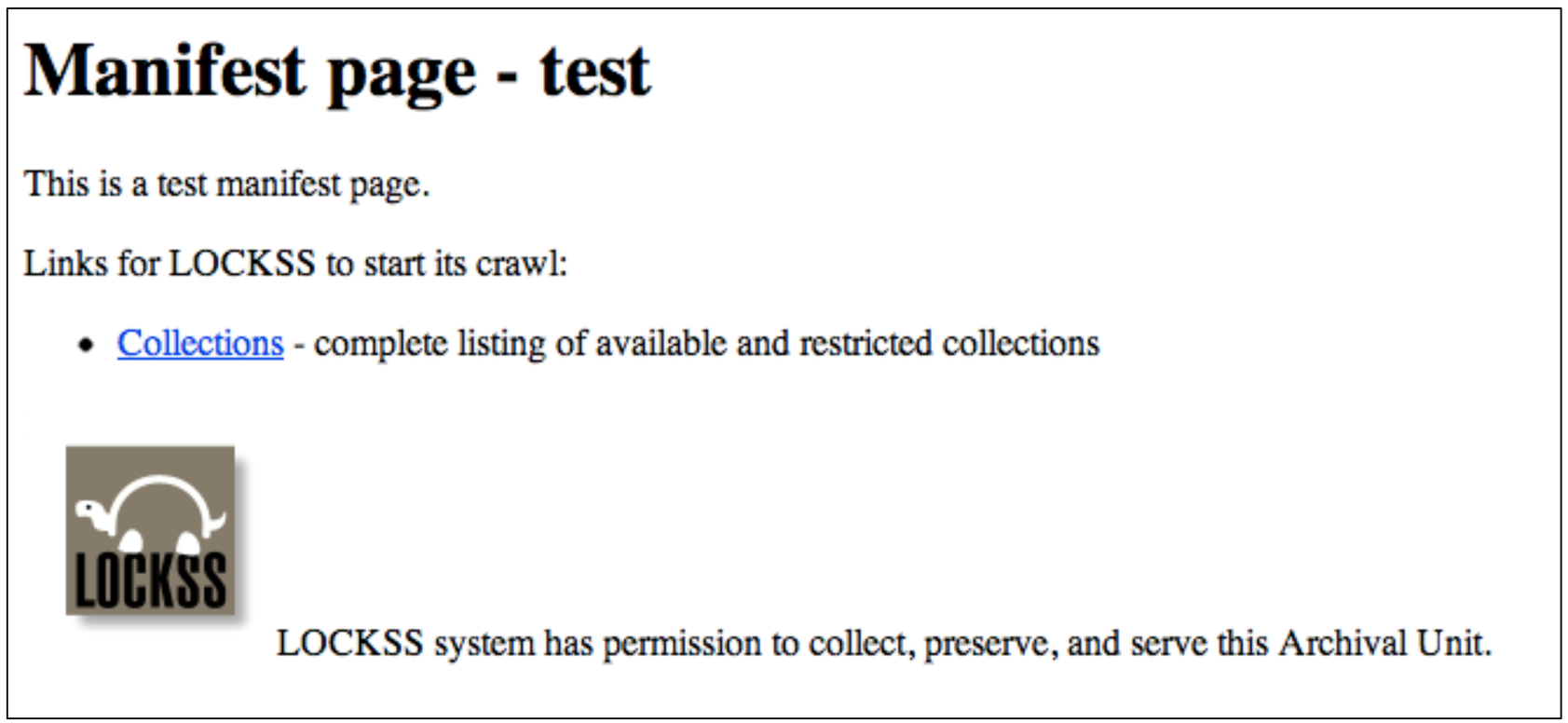
LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)

Every page that needs to be collected by the LOCKSS network shall mention:



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

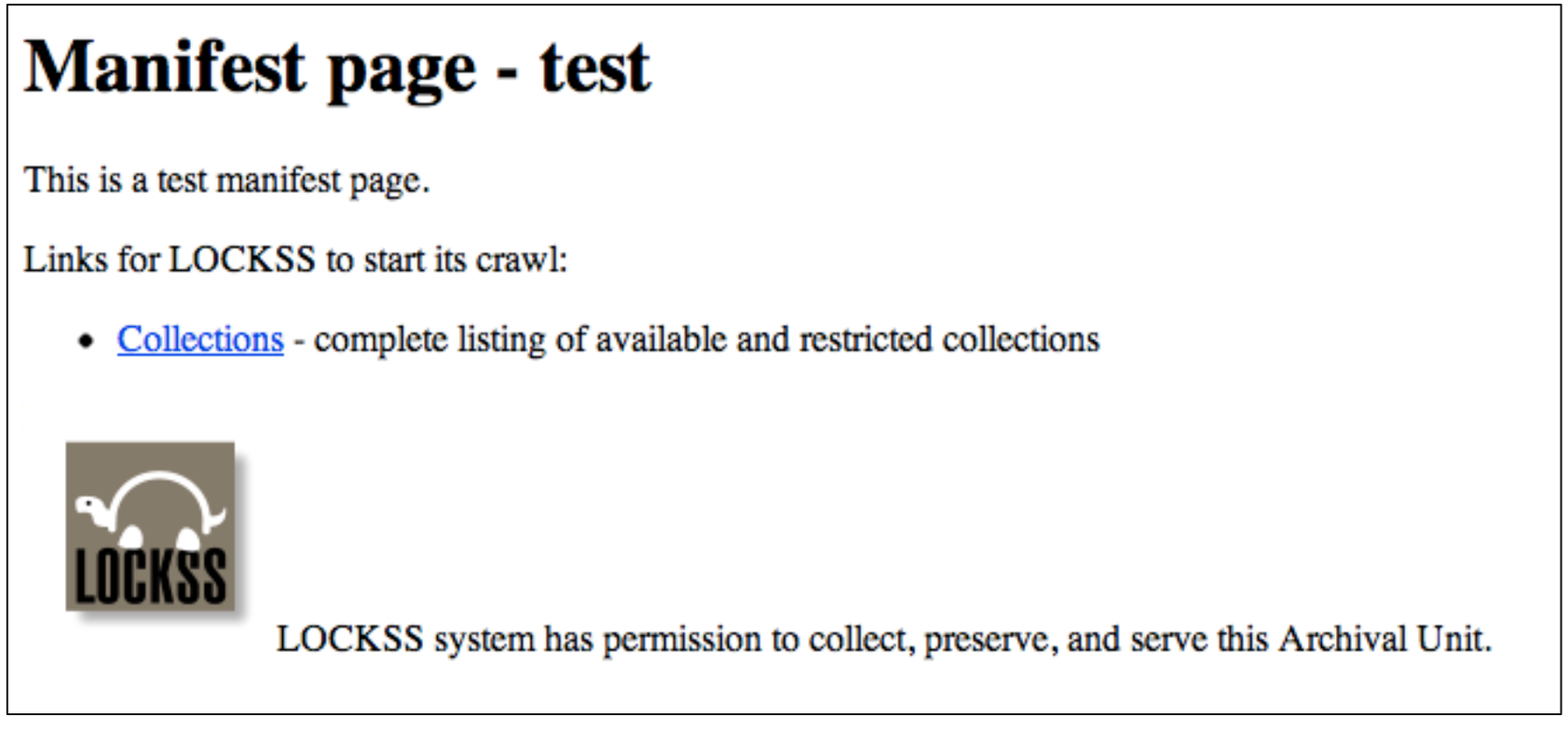
To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)

Every page that needs to be collected by the LOCKSS network shall mention:

```
<p><img align=middle alt="LOCKSS logo" src="http://www.lockss.org/LOCKSS.logo.1%27.gif"
width=80>LOCKSS system has permission to collect, preserve, and serve this Archival Unit.</p>
```



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

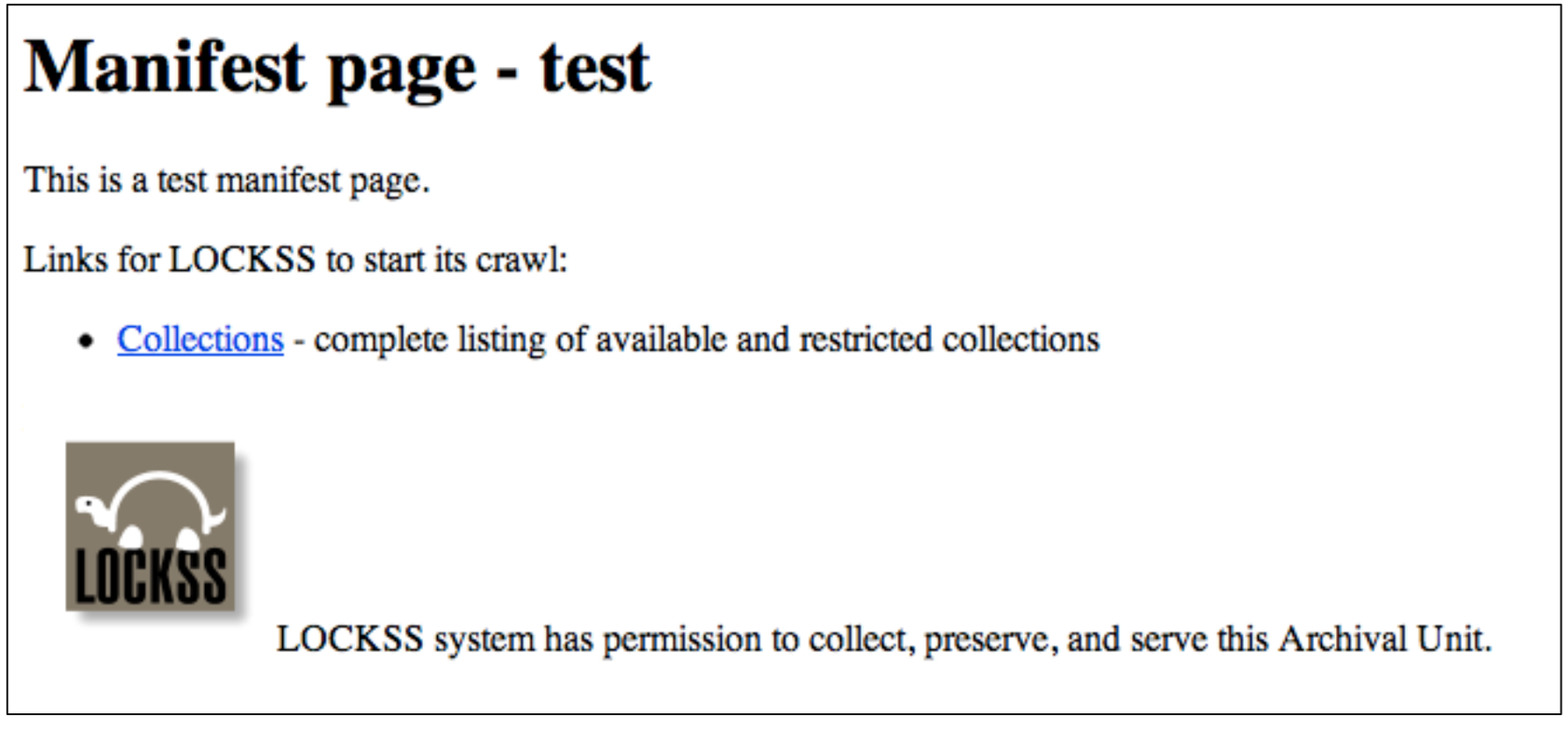
To expose digital collections to your LOCKSS network, you can create a staging server to organize your archives.

Example of website structure:

./lockss/manifest.html (main manifest page for all collections)
./lockss/openaccess/openaccess.html (manifest page for the open-access collection)
./lockss/openaccess/2009/index.html (manifest page for the 2009 open-access collection)

Every page that needs to be collected by the LOCKSS network shall mention:

```
<p><img align=middle alt="LOCKSS logo" src="http://www.lockss.org/LOCKSS.logo.1%27.gif"
width=80>LOCKSS system has permission to collect, preserve, and serve this Archival Unit.</p>
```



Manifest page - test

This is a test manifest page.

Links for LOCKSS to start its crawl:

- Collections - complete listing of available and restricted collections

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

You now created your very own LOCKSS network

http://difusion.academiewb.be/

The future is bright for LOCKSS

https://commons.wikimedia.org/wiki/File:Smiling_turtle.jpg

The current LOCKSS daemon architecture was developed at a time when :

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

monolithic structure

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

monolithic structure

LCAP protocol buried deep into the daemon

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

monolithic structure

LCAP protocol buried deep into the daemon

support local, nearby storage system

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

monolithic structure

LCAP protocol buried deep into the daemon

support local, nearby storage system

AU: not a web archiving standard

The current LOCKSS daemon architecture was developed at a time when :

Only the Global LOCKSS network existed

the web was static

monolithic structure

LCAP protocol buried deep into the daemon

support local, nearby storage system

AU: not a web archiving standard

one box = one machine

New architecture:
LOCKSS As A WebService (LAAWS)

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

independent modules interacting through REST-API

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

independent modules interacting through REST-API

de-silo components: make LCAP an independent module

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

independent modules interacting through REST-API

de-silo components: make LCAP an independent module

support for large-scale distributed storage

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

independent modules interacting through REST-API

de-silo components: make LCAP an independent module

support for large-scale distributed storage

align with web archiving standards (WARC-native)

New architecture: LOCKSS As A WebService (LAAWS)

profit from web-scale OSS to cope with fast evolving web

independent modules interacting through REST-API

de-silo components: make LCAP an independent module

support for large-scale distributed storage

align with web archiving standards (WARC-native)

one box = multiple containers possibly on different machines

Agenda

Motivation

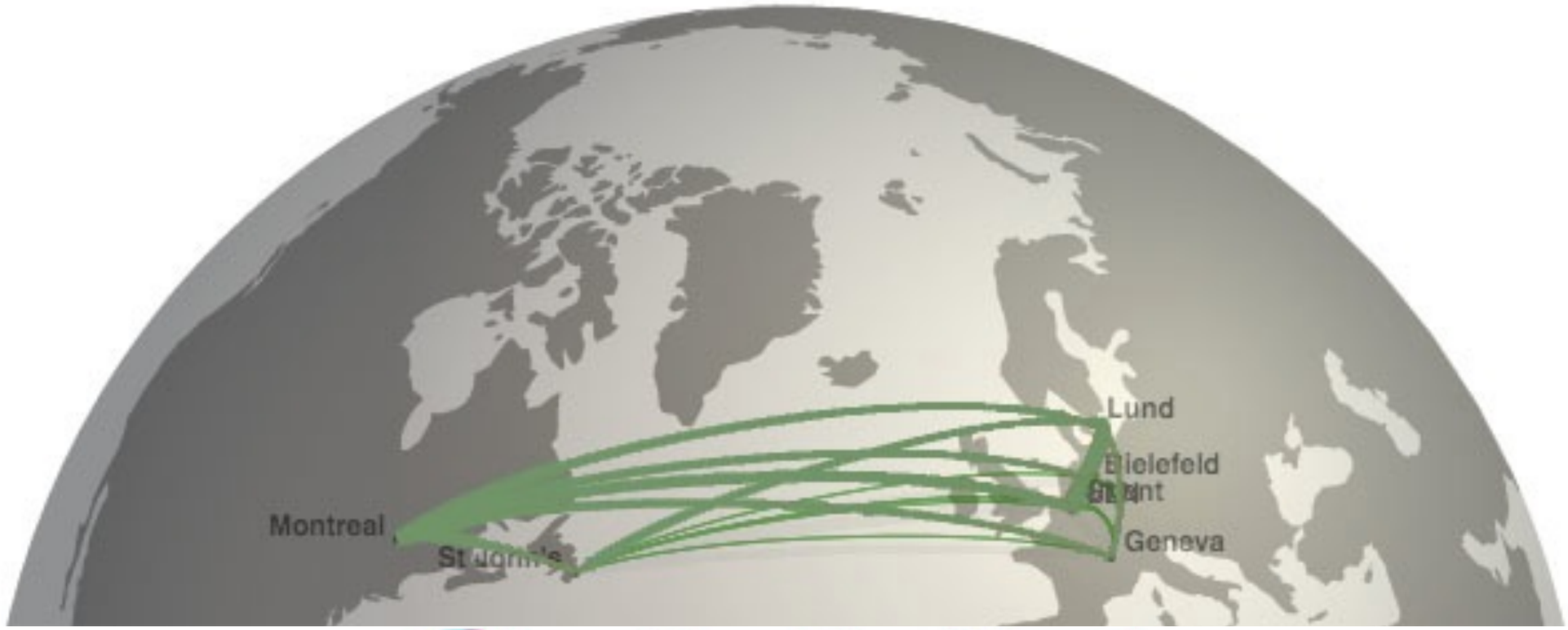
LOCKSS main concepts

SAFE PLN

Build your own LOCKSS Network

Discussions

LEARNING RESOURCES



SAFE

SAFE Archiving FEderation