Reference Rot in Scholarly Communication: A Reliable Quantification and a Proposed Solution

Martin Klein
@mart1nkle1n

Research Library
Los Alamos National Laboratory

Acknowledgements:
Herbert Van de Sompel, Shawn Jones, Harihar Shankar (LANL)
Richard Tobin, Claire Grover (University of of Edinburgh)
Andy Jackson (British Library)

CC BY

Los Alamos NATIONAL LABORATORY

Agenda

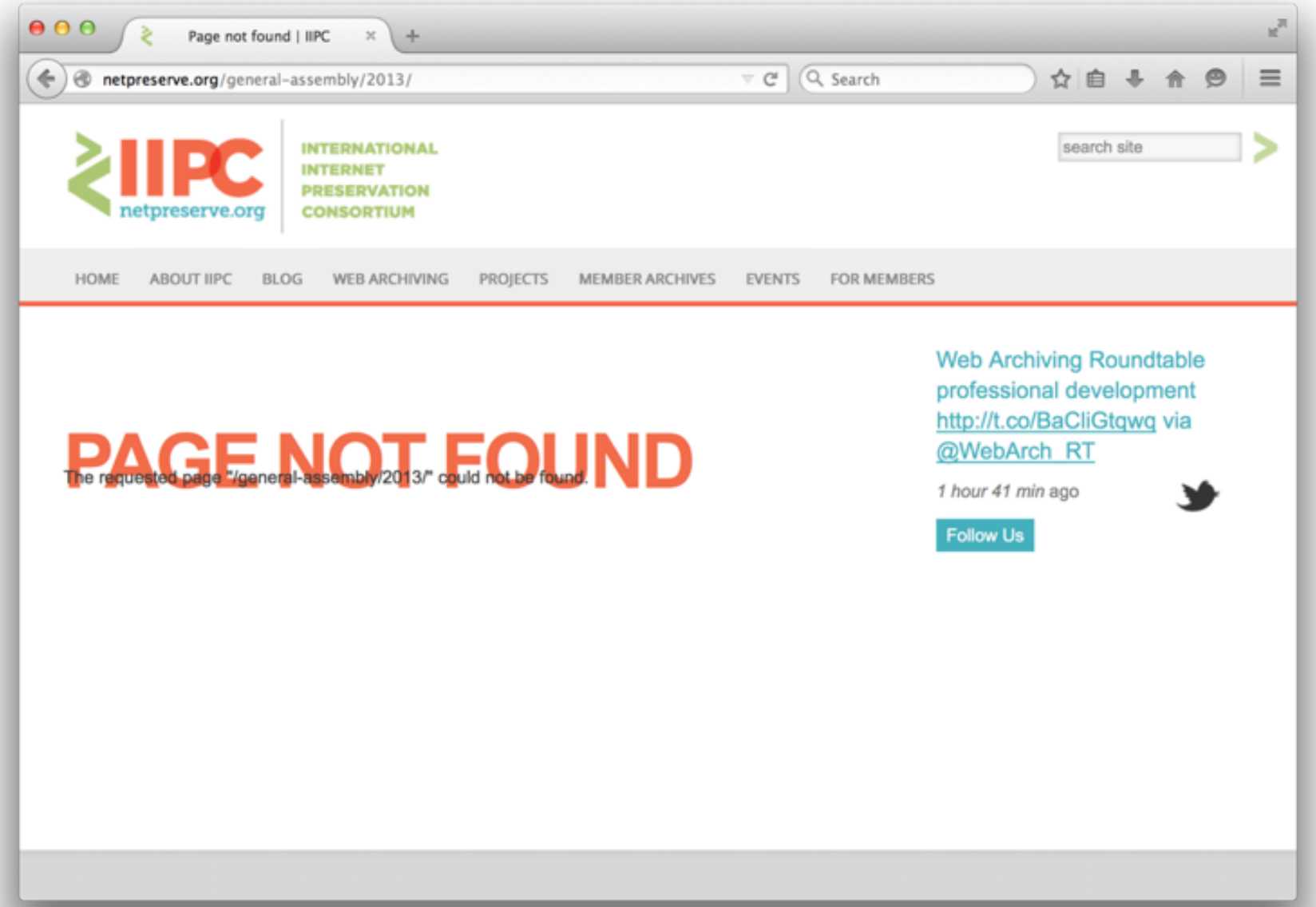
1. Definition: Reference Rot
2. Study: Quantification of Content Drift
3. Proposed Solution: Robust Links

CC BY

Agenda

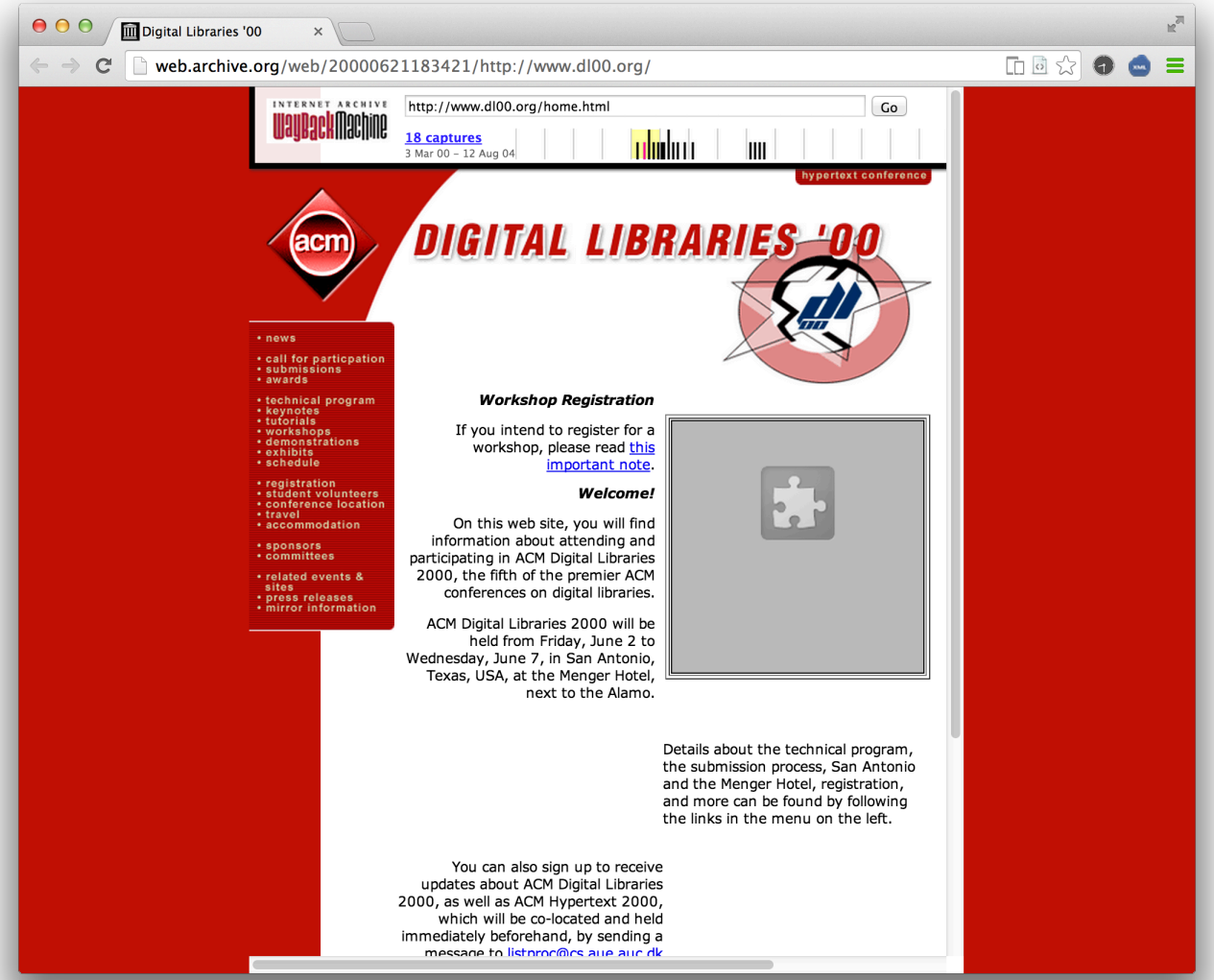
1. Definition: Reference Rot
2. Study: Quantification of Content Drift
3. Proposed Solution: Robust Links

CC BY



CC BY

Los Alamos NATIONAL LABORATORY

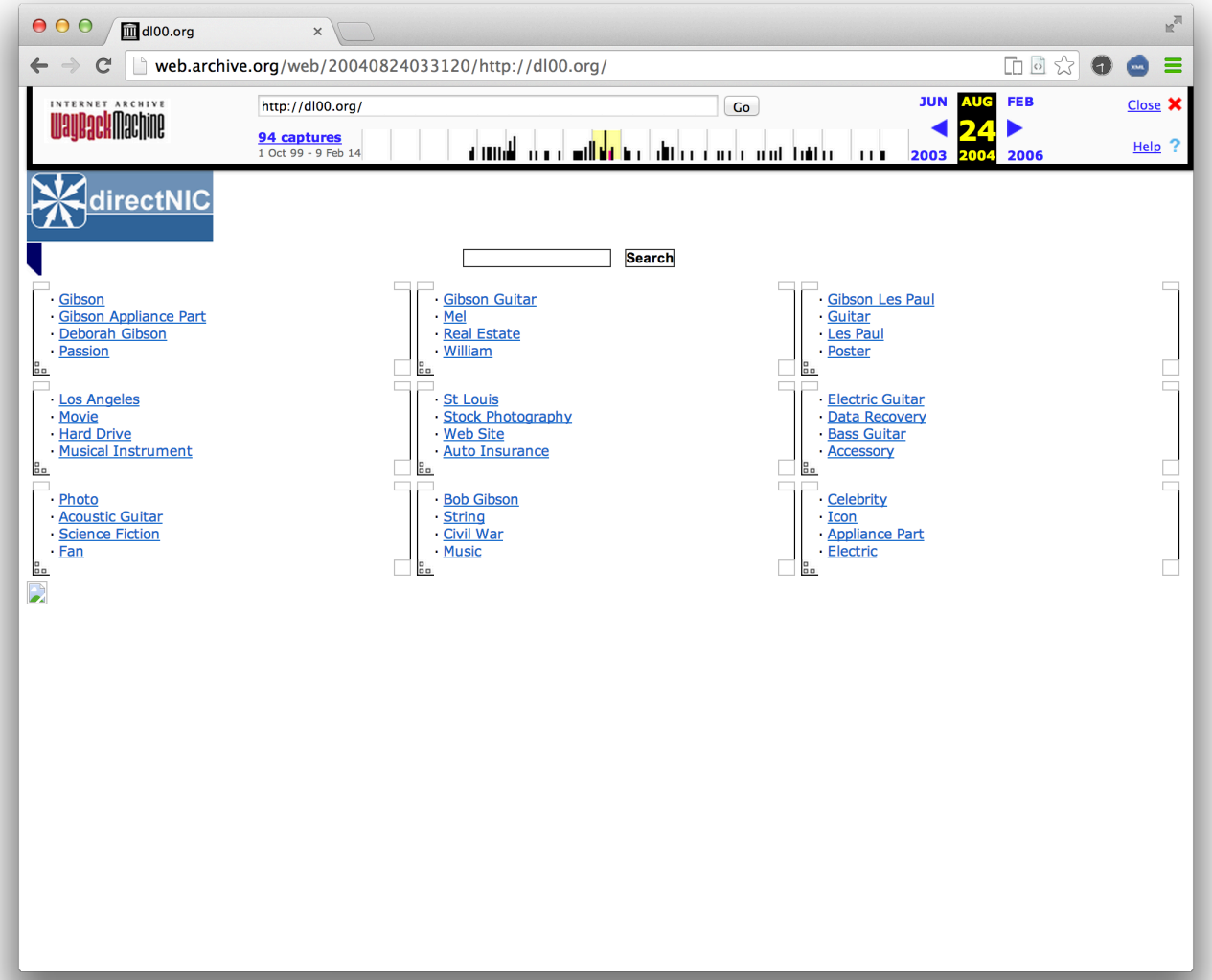


http://dl00.org

2000

CC BY

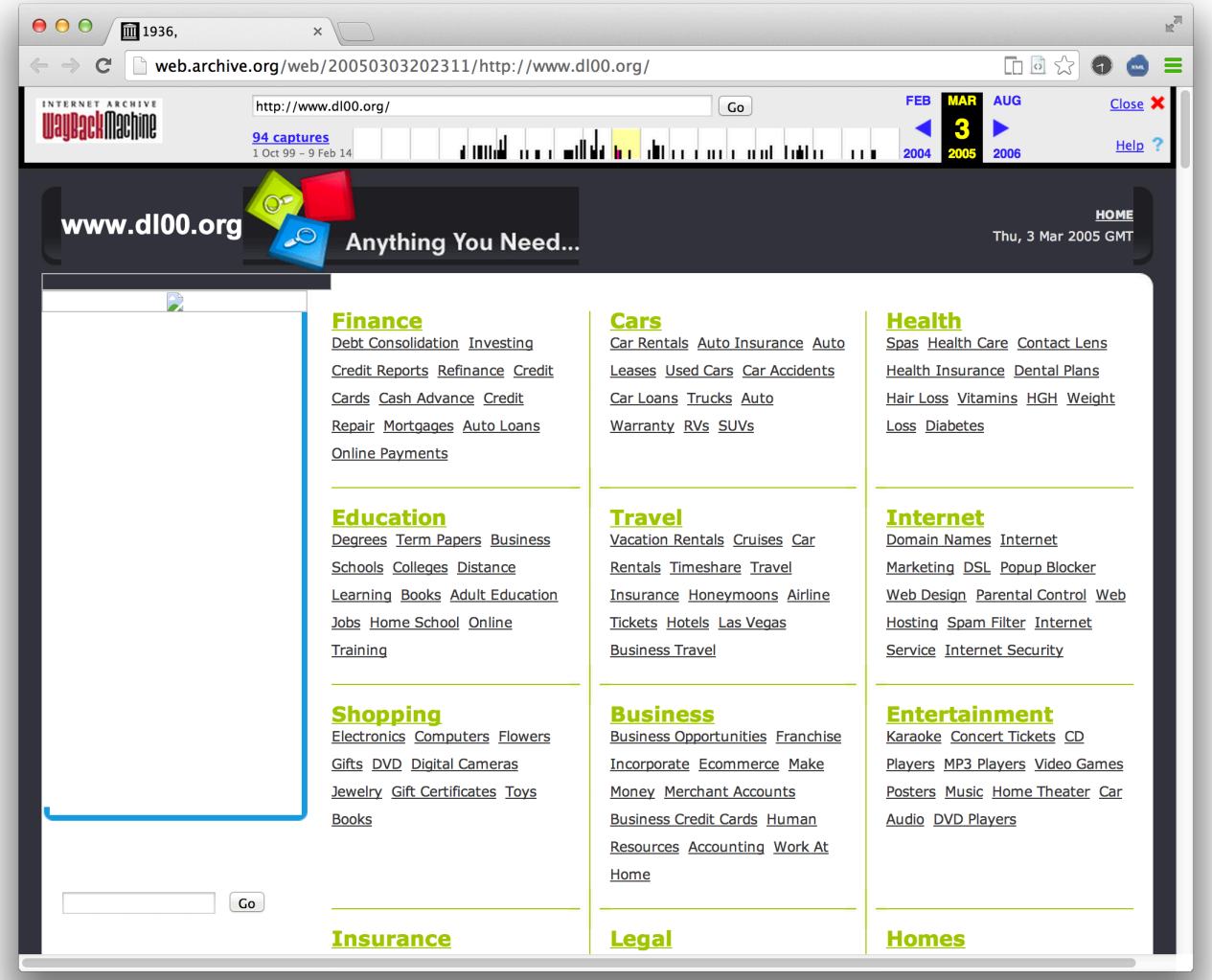
Los Alamos NATIONAL LABORATORY



http://dl00.org

2004

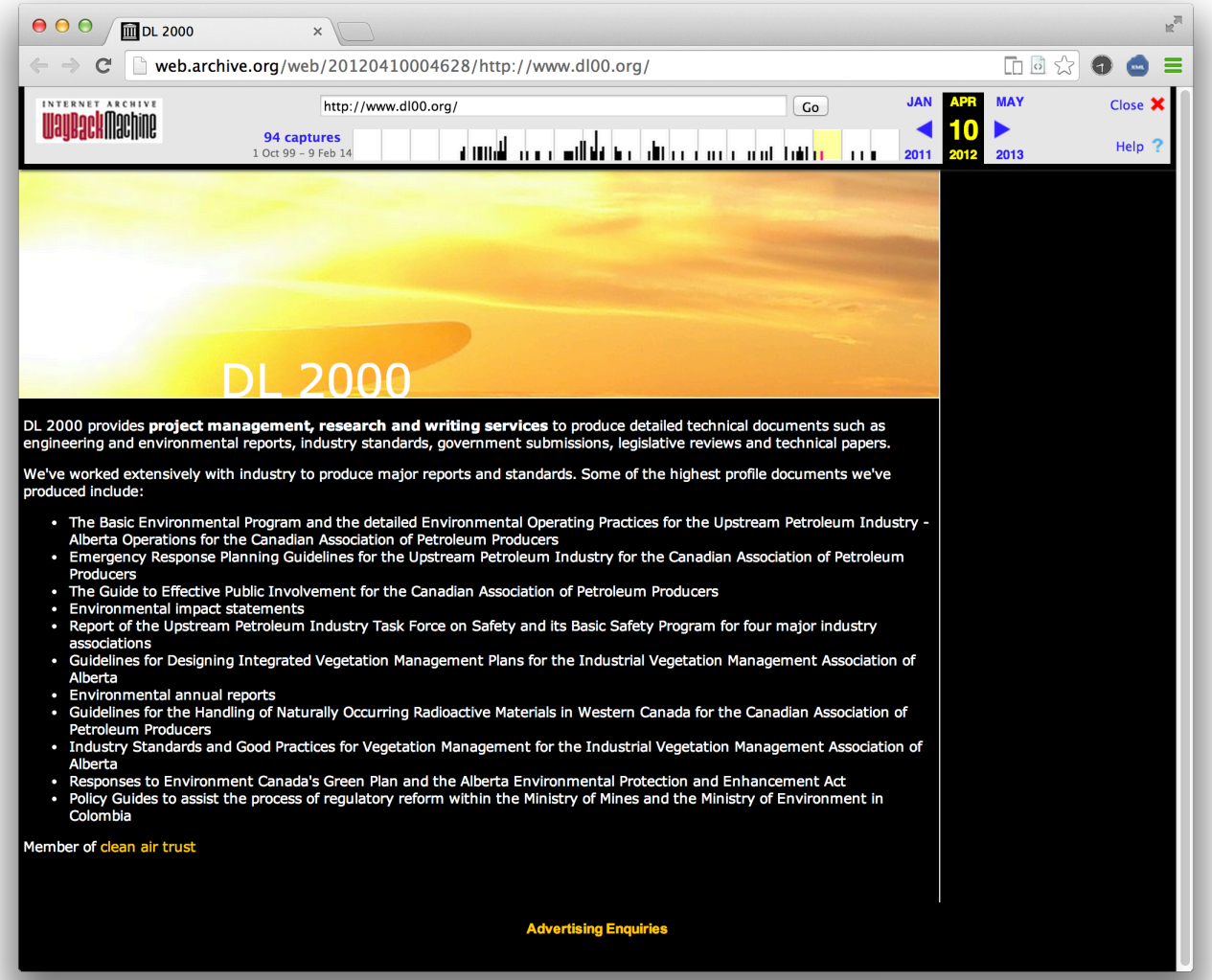
CC BY



http://dl00.org

2005

CC BY



http://dl00.org

2008

CC BY

Los Alamos NATIONAL LABORATORY

hiberlink
http://hiberlink.org/

Definition:

- Link Rot + Content Drift = Reference Rot

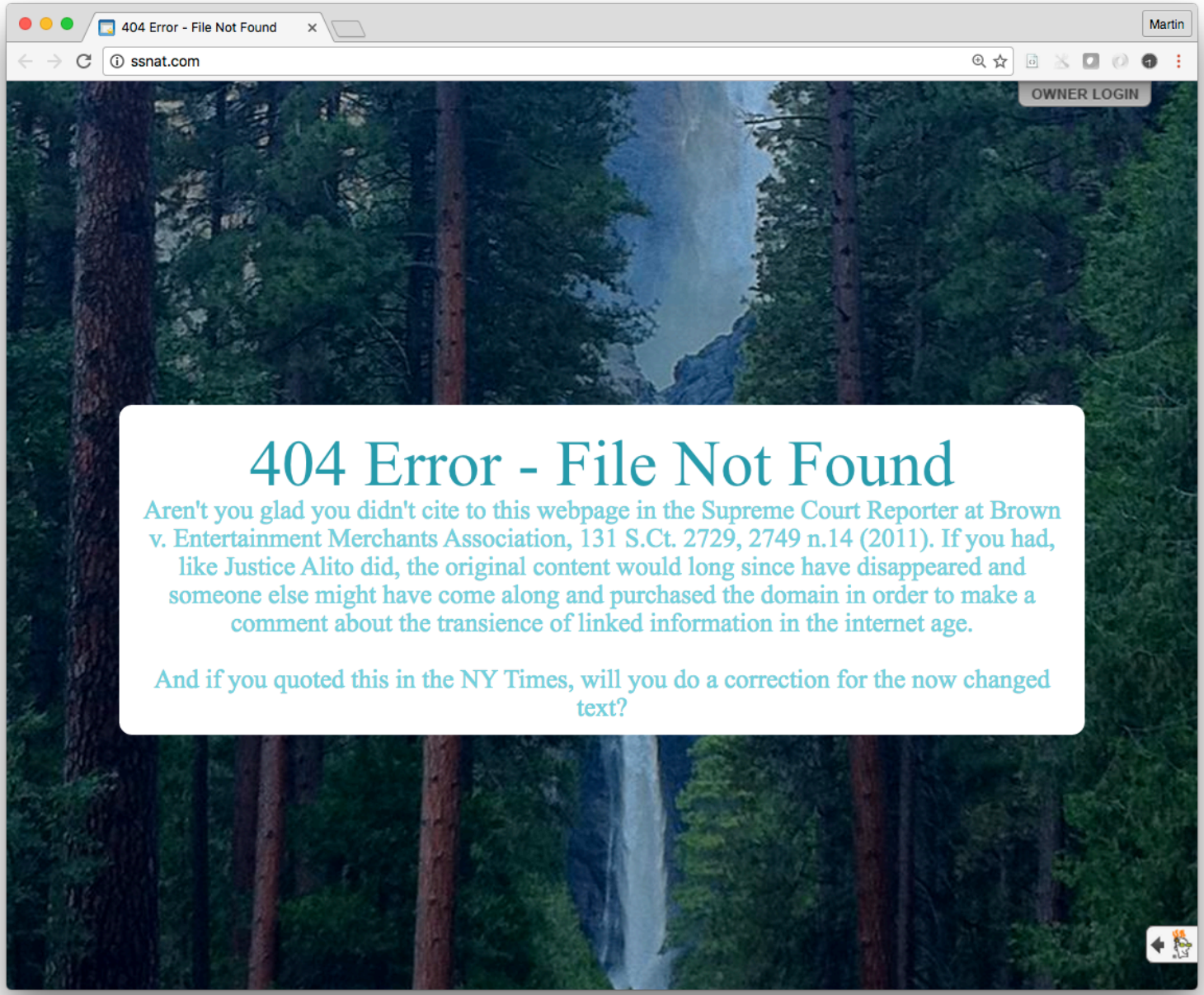
Observation:

- Links to these resources are subject to Reference Rot
- Web at large resources referenced in scholarly articles

Problem:

- Thread to integrity of the web-based scholarly record
- Resources do not have the same sense of fixity like e.g., journal articles
- Resources' custodianship is different, in terms of long-term archiving, integrity, and access

CC BY


404 Error - File Not Found
Martin
ssnat.com
OWNER LOGIN
404 Error - File Not Found
Aren't you glad you didn't cite to this webpage in the Supreme Court Reporter at Brown v. Entertainment Merchants Association, 131 S.Ct. 2729, 2749 n.14 (2011). If you had, like Justice Alito did, the original content would long since have disappeared and someone else might have come along and purchased the domain in order to make a comment about the transience of linked information in the internet age.
And if you quoted this in the NY Times, will you do a correction for the now changed text?

CC BY

plos.org create account sign in

PLOS ONE | Publish About Browse Search

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

Published: December 26, 2014 • http://dx.doi.org/10.1371/journal.pone.0115253

71 Save | 12 Citation | 18,506 View | 332 Share

Article | Authors | Metrics | Comments | Related Content

Download PDF | Print | Share

Abstract
Introduction
Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Reader Comments (0)
Media Coverage (2)
Figures

Abstract

The emergence of the web has fundamentally affected most aspects of information communication, including scholarly communication. The immediacy that characterizes publishing information to the web, as well as accessing it, allows for a dramatic increase in the speed of dissemination of scholarly knowledge. But, the transition from a paper-based to a web-based scholarly communication system also poses challenges. In this paper, we focus on reference rot, the combination of link rot and content drift to which references to web resources included in Science, Technology, and Medicine (STM) articles are subject. We investigate the extent to which reference rot impacts the ability to revisit the web context that surrounds STM articles some time after their publication. We do so on the basis of a vast collection of articles from three corpora that span publication years 1997 to 2012. For over one million references to web resources extracted from over 3.5 million articles, we determine whether the HTTP URI is still responsive on the live web and whether web archives contain an archived snapshot representative of the state the referenced resource had at the time it was referenced. We observe that the fraction of articles containing references to web resources is growing steadily over time. We find one out of five STM articles suffering from reference rot, meaning it is impossible to revisit the web context that surrounds them some time after their publication. When only considering STM articles that contain references to web resources, this fraction increases to seven out of ten. We suggest that, in order to safeguard the long-term integrity of the web-based scholarly record, robust solutions to combat the reference rot problem are required. In conclusion, we provide a brief insight into the directions that are explored with this regard in the context of the Hiberlink project.

CrossMark

Subject Areas

Archives
Internet
Extrapolation
Communications
Computer and infor...
Evolutionary immun...
Species interactions
Publication ethics

ADVERTISEMENT

http://dx.doi.org/10.1371/journal.pone.0115253

CC BY

Los Alamos NATIONAL LABORATORY

Agenda

1. Definition: Reference Rot
2. Study: Quantification of Content Drift
3. Proposed Solution: Robust Links

CC BY

plos.org create account sign in

PLOS ONE | Publish About Browse Search

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content

Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, Claire Grover

Published: December 2, 2016 • http://dx.doi.org/10.1371/journal.pone.0167475

1 Save | 0 Citation | 482 View | 42 Share

Article | Authors | Metrics | Comments | Related Content

Download PDF | Print | Share

Abstract
Introduction
Related Work
Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Reader Comments (0)
Media Coverage (0)
Figures

Abstract

Increasingly, scholarly articles contain URI references to "web at large" resources including project web sites, scholarly wikis, ontologies, online debates, presentations, blogs, and videos. Authors reference such resources to provide essential context for the research they report on. A reader who visits a web at large resource by following a URI reference in an article, some time after its publication, is led to believe that the resource's content is representative of what the author originally referenced. However, due to the dynamic nature of the web, that may very well not be the case. We reuse a dataset from a previous study in which several authors of this paper were involved, and investigate to what extent the textual content of web at large resources referenced in a vast collection of Science, Technology, and Medicine (STM) articles published between 1997 and 2012 has remained stable since the publication of the referencing article. We do so in a two-step approach that relies on various well-established similarity measures to compare textual content. In a first step, we use 19 web archives to find snapshots of referenced web at large resources that have textual content that is representative of the state of the resource around the time of publication of the referencing paper. We find that representative snapshots exist for about 30% of all URI references. In a second step, we compare the textual content of representative snapshots with that of their live web counterparts. We find that for over 75% of references the content has drifted away from what it was when referenced. These results raise significant concerns regarding the long term integrity of the web-based scholarly record and call for the deployment of techniques to combat these problems.

CrossMark

Subject Areas

Archives
Birds
Internet
Cats
Algorithms
Crawling
Ontologies
Species interactions

ADVERTISEMENT

http://dx.doi.org/10.1371/journal.pone.0167475

CC BY

Study Dataset

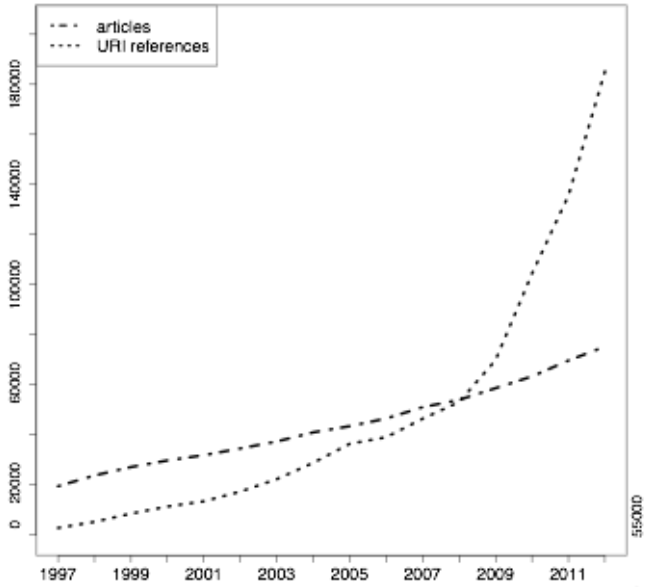
- 3.5 million articles from arXiv, Elsevier, PMC
- Published between Jan 1997 – Dec 2012
- Converted from PDF to XML
- Extraction of URIs to web at large resources (>1 million)
- Keep track of articles' publication date

	arXiv	**Elsevier**	**PMC**
total articles	707,667	2,285,000	595,889
articles with HTTP references	142,134	94,645	156,160
amount of HTTP references	346,177	232,712	480,853

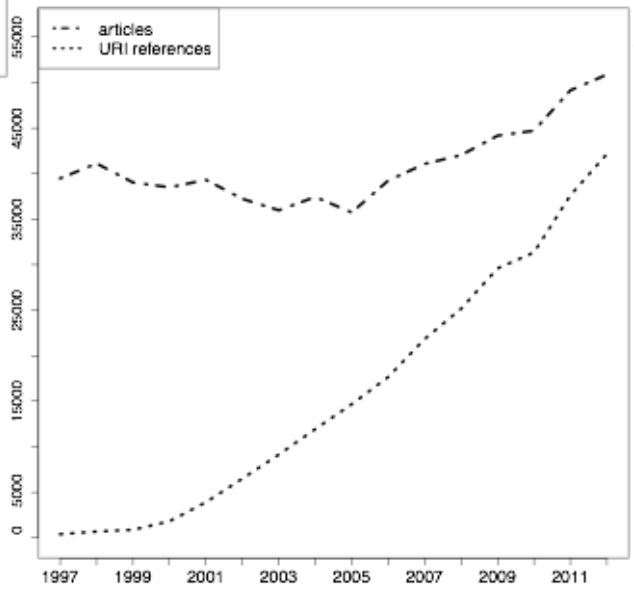
CC BY

Corpora

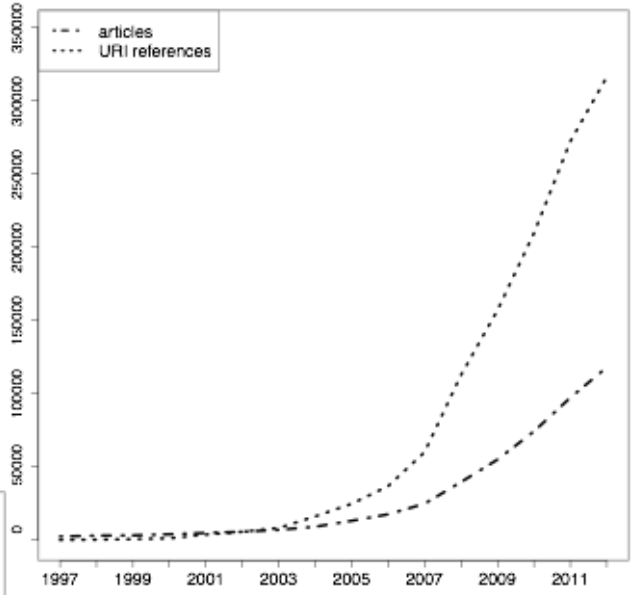
arXiv



Elsevier



PMC



CC BY

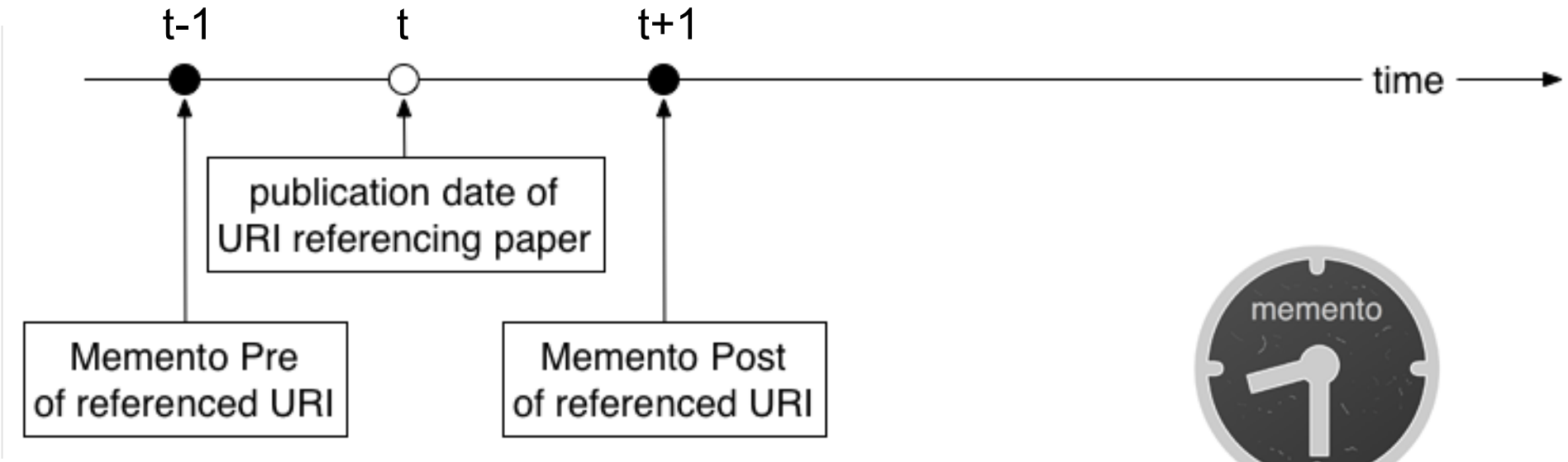
Los Alamos NATIONAL LABORATORY

Novel Approach to Assess Content Drift

CC BY

Step 1: Find Mementos

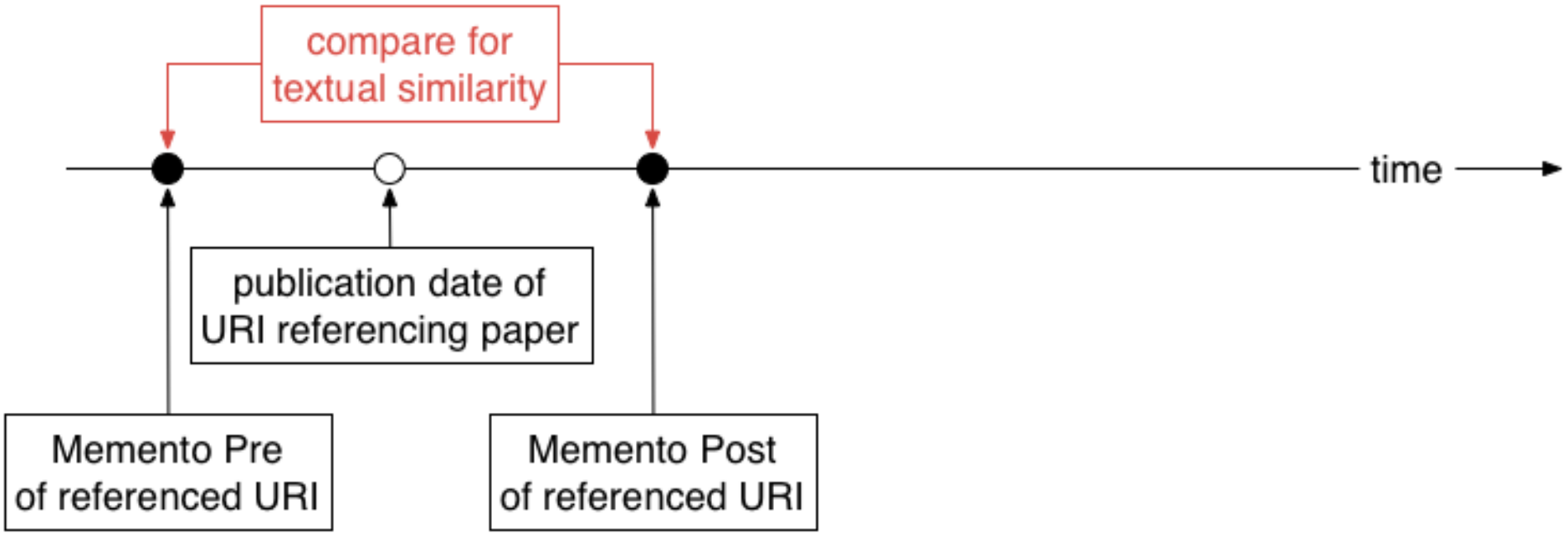
- ~ 1 million URI references
- ~ 650k Memento Pre/Post pairs discovered via Memento



https://mementoweb.org
https://tools.ietf.org/html/rfc7089

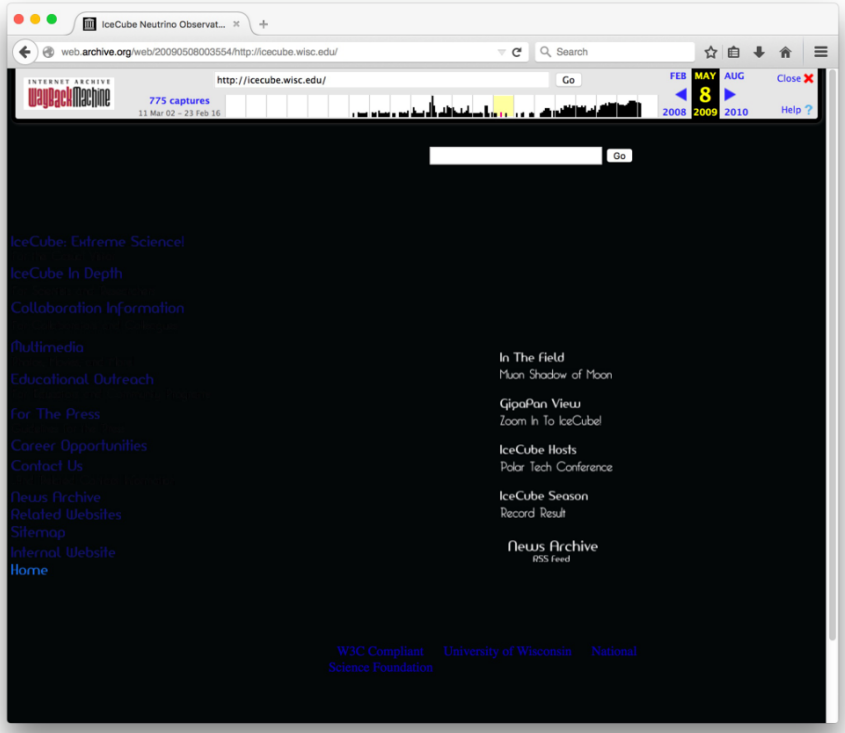
CC BY

Step 2: Select Representative Mementos

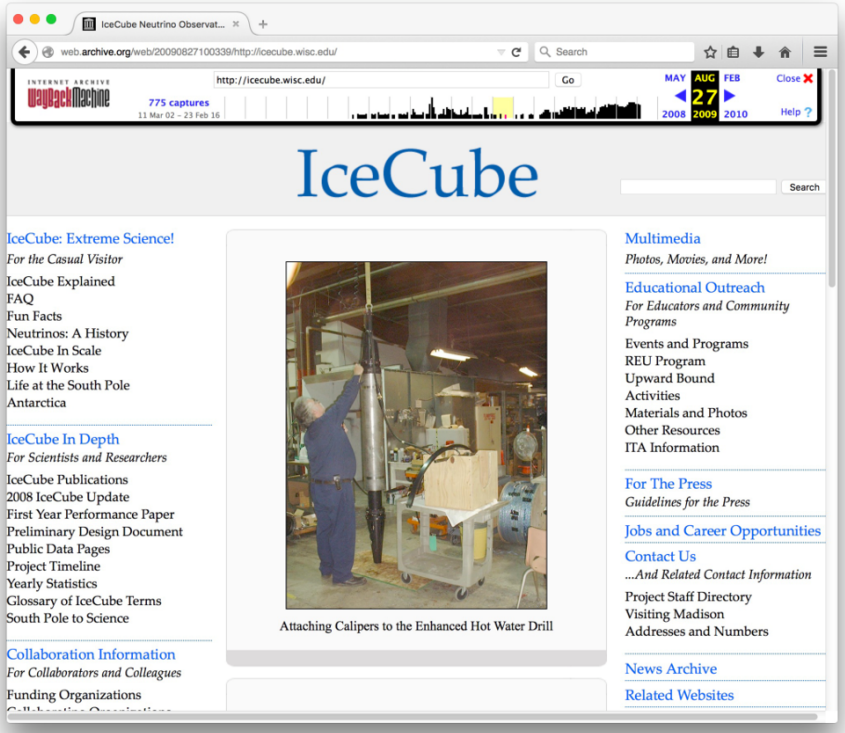


CC BY

Referenced in
http://dx.doi.org/10.1016/j.nuclphysa.2009.05.110
published on August 15th 2009



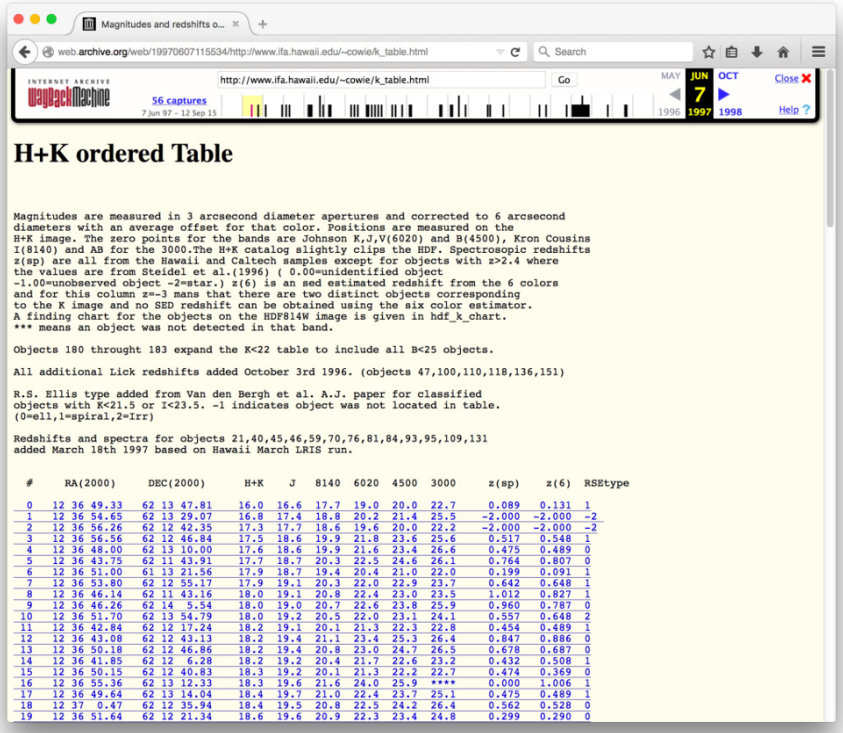
May 8th 2009



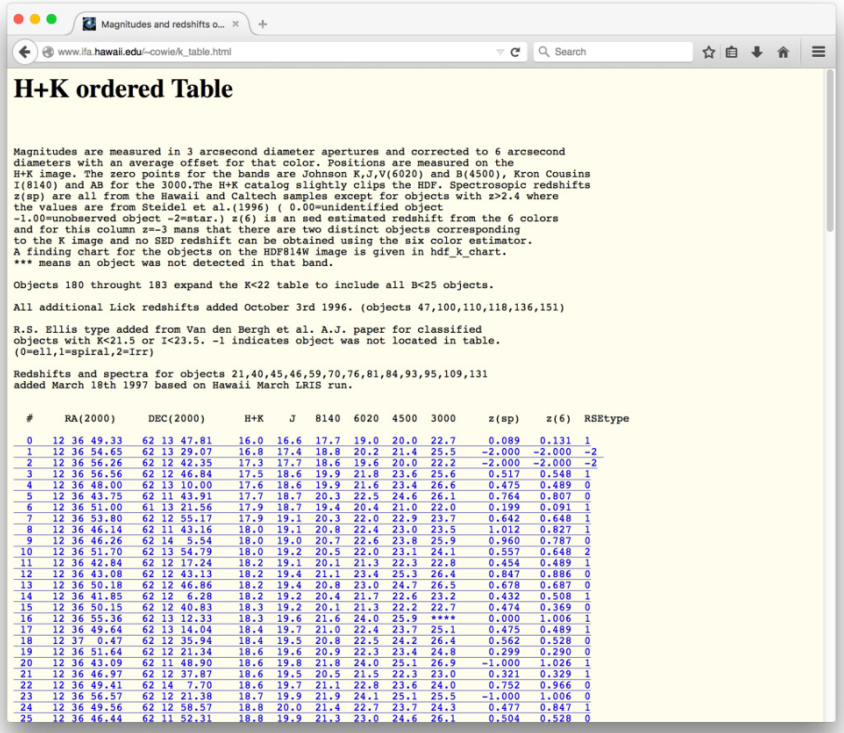
August 27th 2009

CC BY

Referenced in
http://arxiv.org/abs/astro-ph/9707064
published on July 4th 1977



June 7th 1997

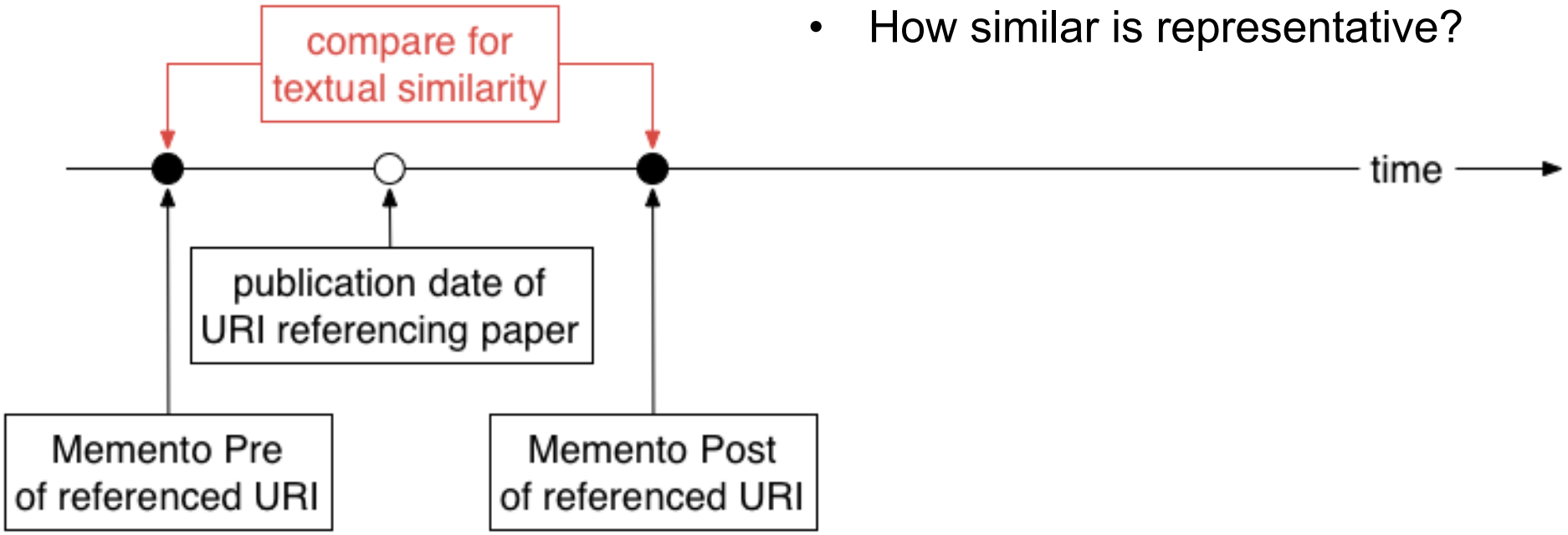


today

CC BY

Step 2: Select Representative Mementos

- Apply content similarity measures
- How similar is representative?



CC BY

Content Similarity Measures

- Compute normalized scores (values between 0...100) for:
  - Simhash
  - Jaccard
  - Sørensen-Dice
  - Cosine

$s_1 = "birds\ eat\ cats"$ $s_2 = "cats\ eat\ birds"$ $s_3 = "birds\ eat\ cake"$		
	s_1 vs. s_2	s_1 vs. s_3
Simhash	67.2	73.4
Jaccard	100.0	90.9
Sørensen-Dice	100.0	95.2
Cosine	100.0	50.3

CC BY

Representative Mementos

- Idea
  - If perfect score in all 4 similarity measures

  → Memento Pre and Post are the same

  → Representative Mementos

- Sanity check needed
  - Via HTTP headers: E-Tag and Last-Modified
  - If same for Pre and Post Memento

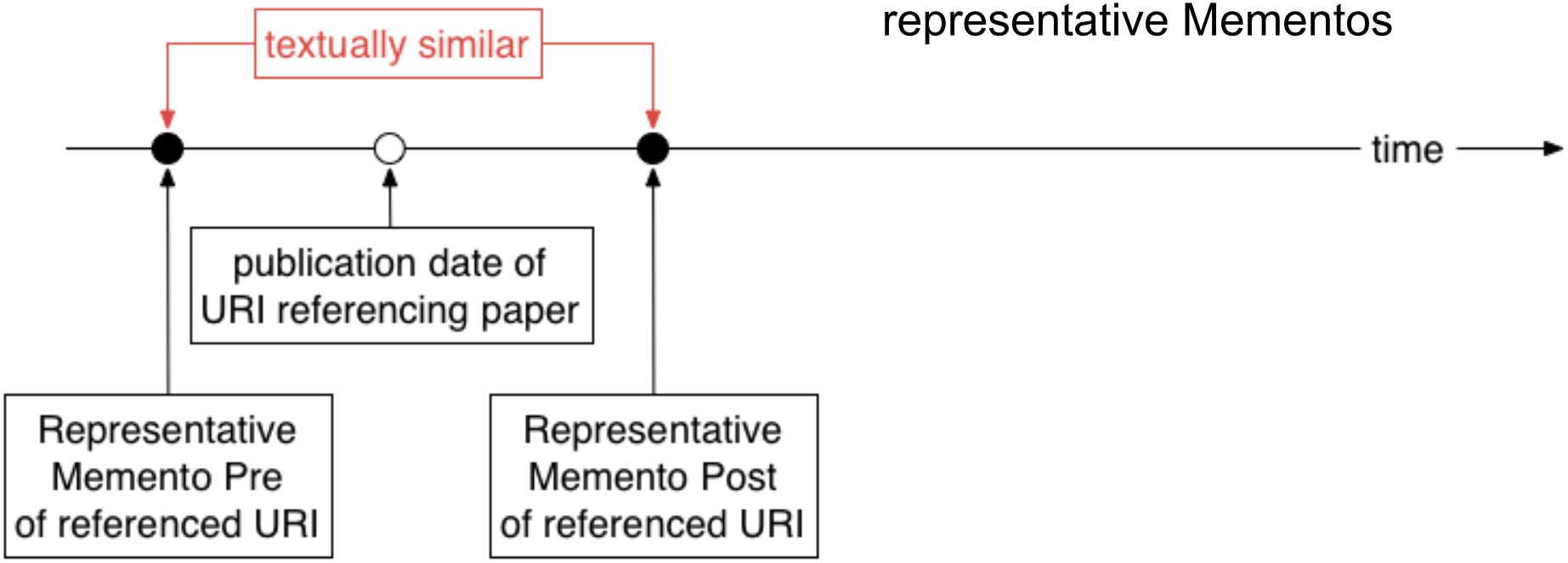
  → HTTP-same

- Sanity check passed!
  - 98.88% of Memento pairs that are HTTP-same have perfect score in all 4 similarity measures

CC BY

Step 2: Select Representative Mementos

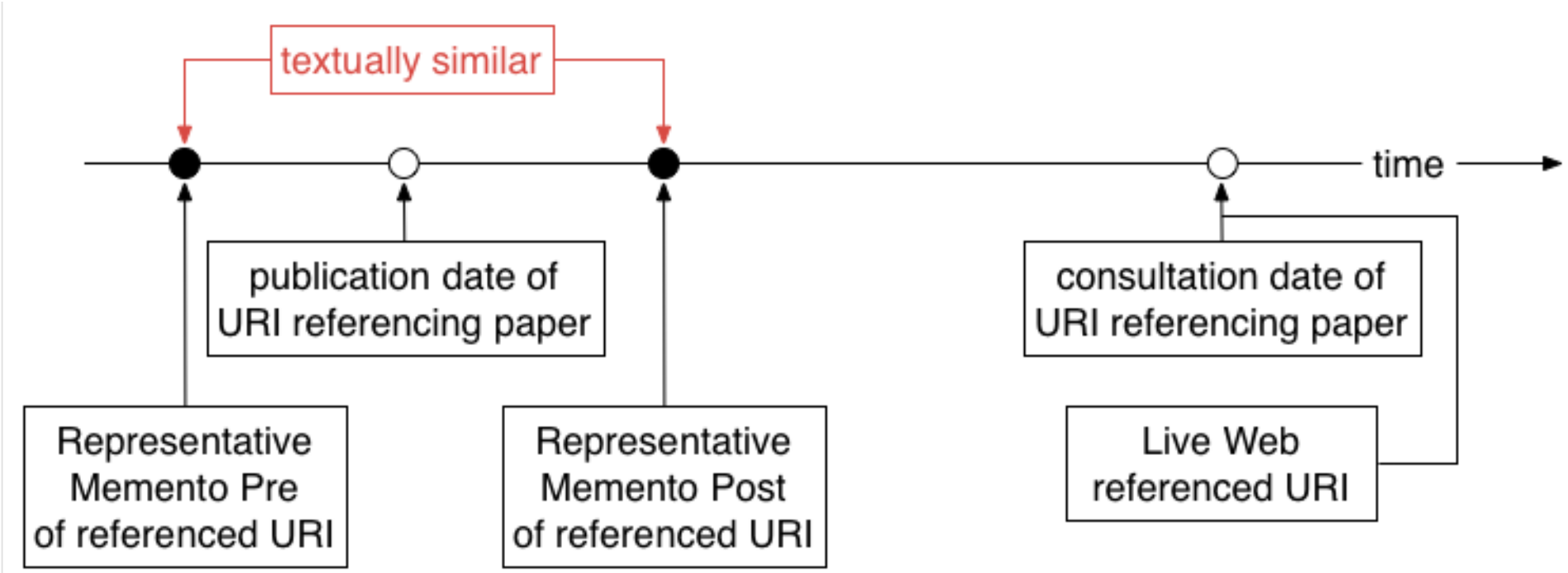
- ~ 313k referenced URIs have representative Mementos



CC BY

Step 3: Dereference Live Web Version of URI

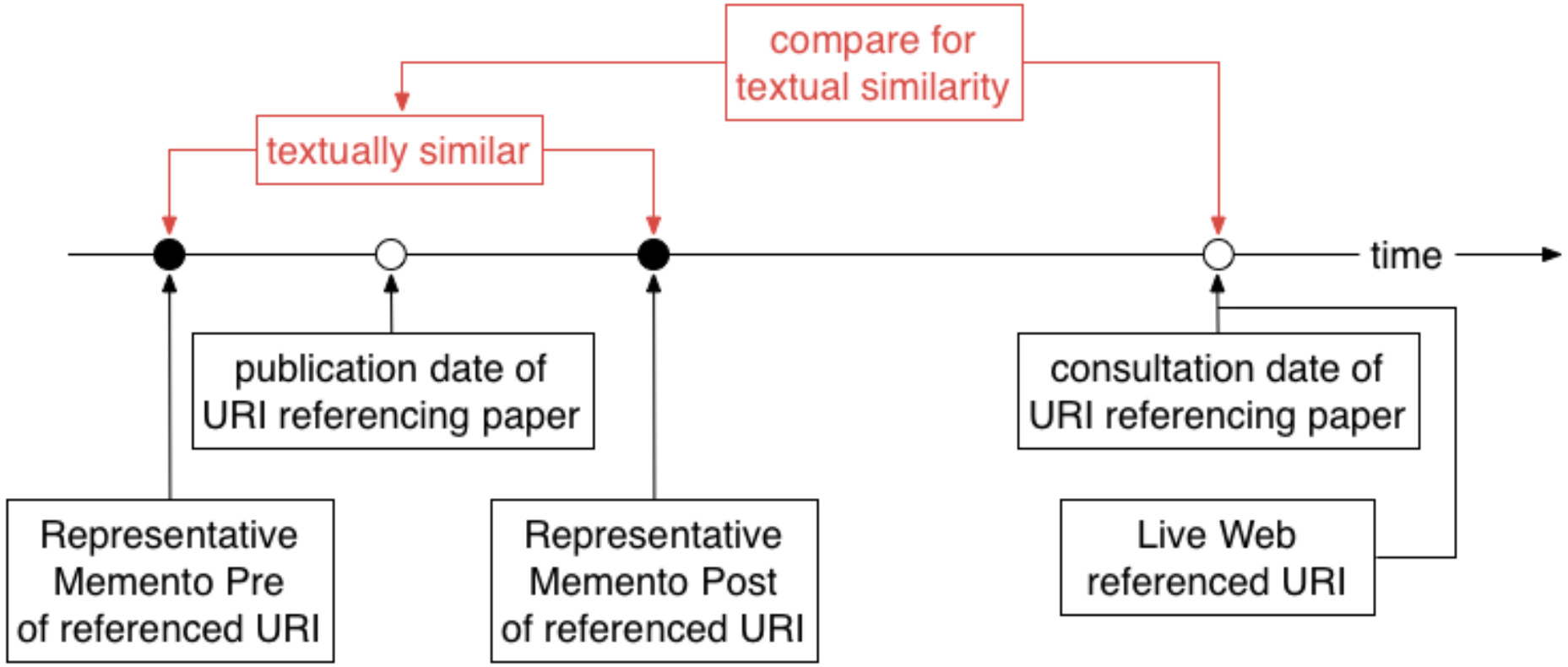
- 241k out of 313k URIs have a live web version



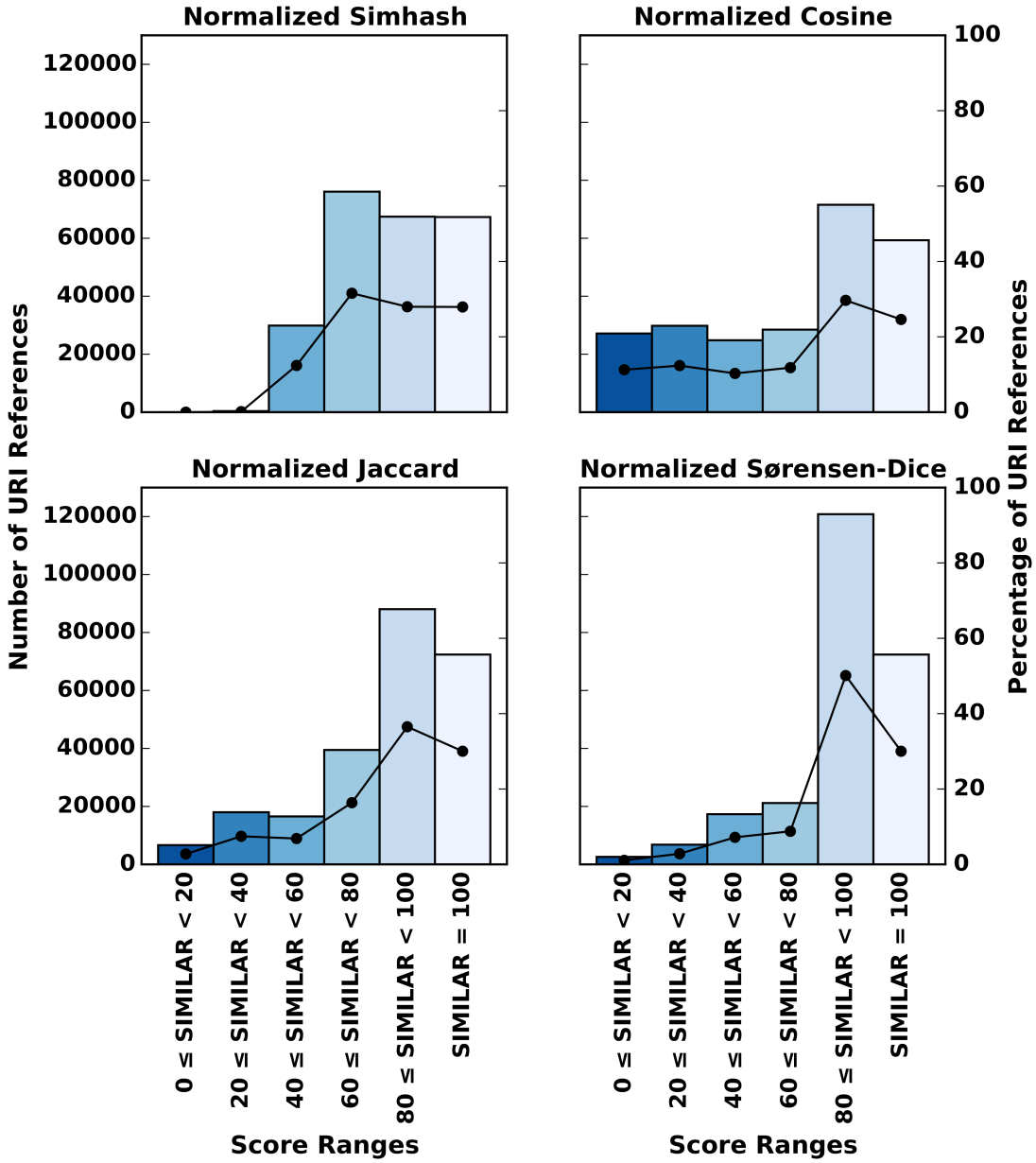
CC BY

Step 4: Representative Memento vs. Live Version

- Apply content similarity measures
- Bin results into 6 clusters



CC BY



CC BY

Los Alamos NATIONAL LABORATORY

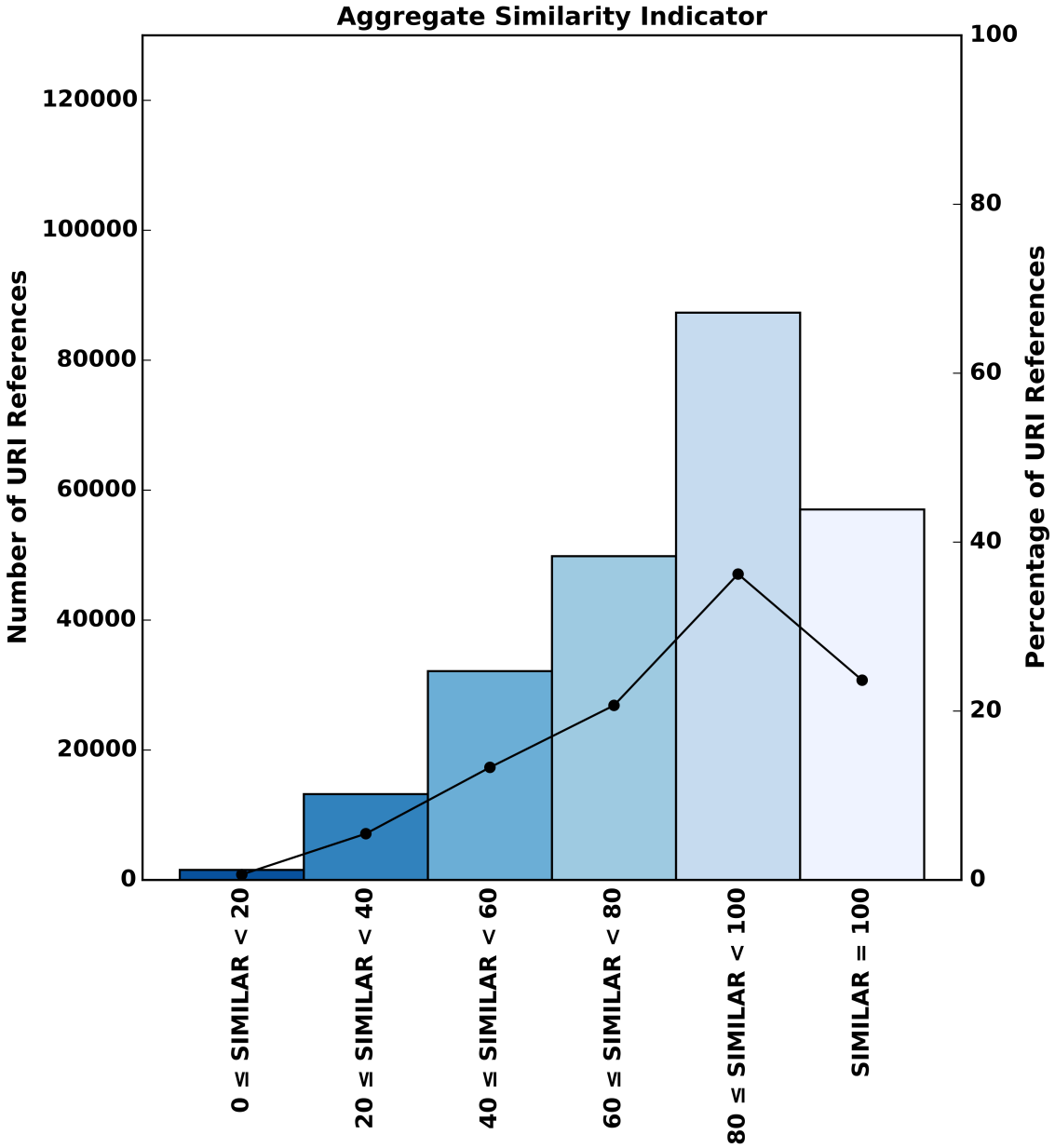
Aggregate Similarity Score

Good:

23.7% of URIs have *not* drifted!

Bad:

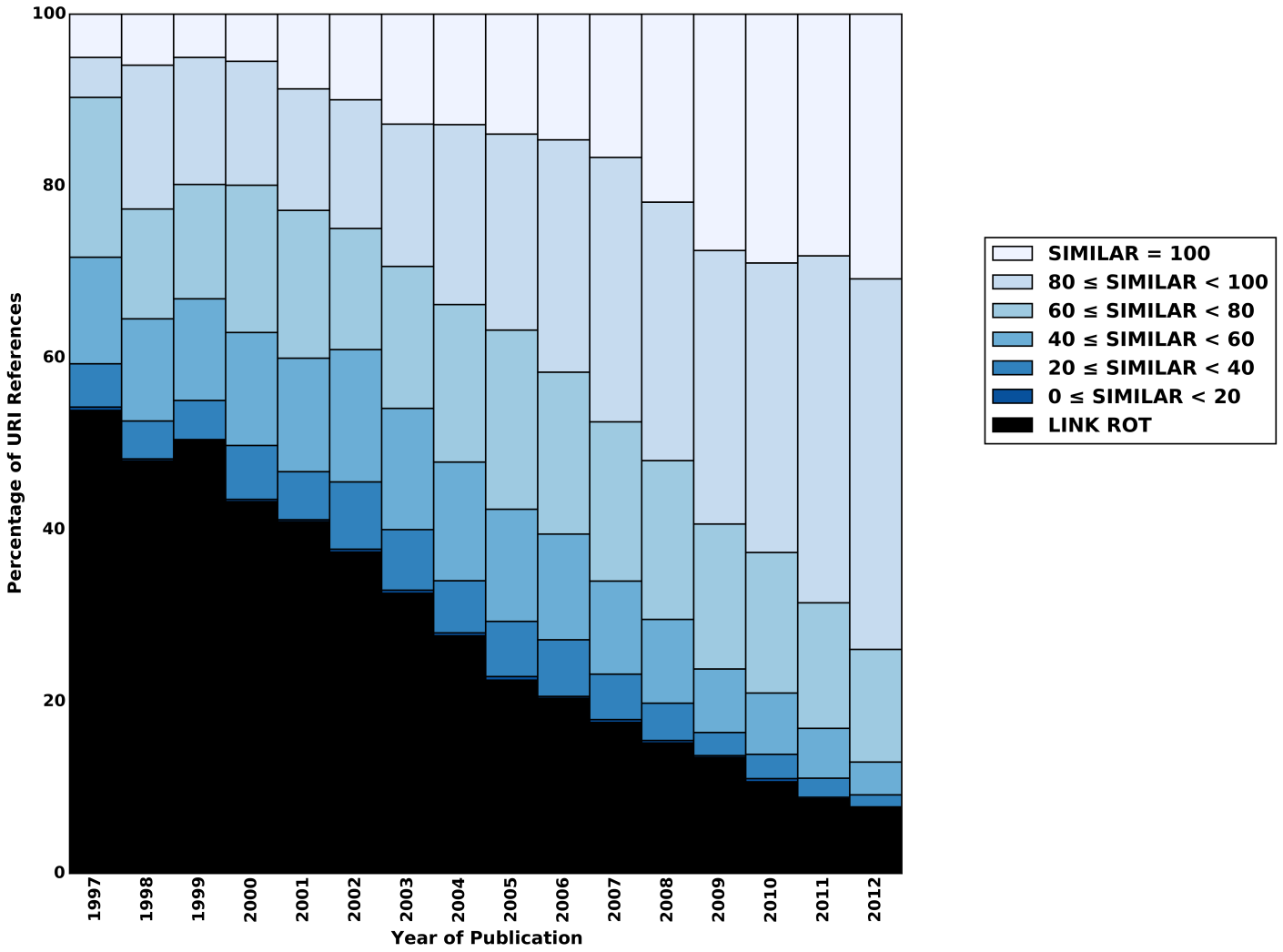
3/4 URIs *have* drifted!



CC BY

Los Alamos NATIONAL LABORATORY

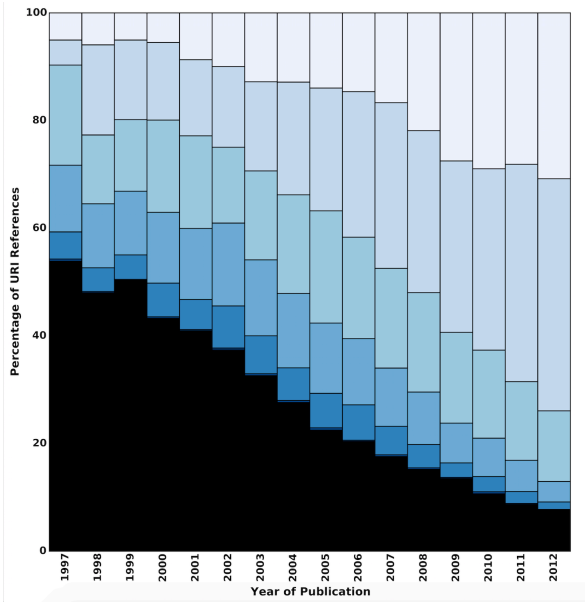
Content Drift & Link Rot Over Time - arXiv



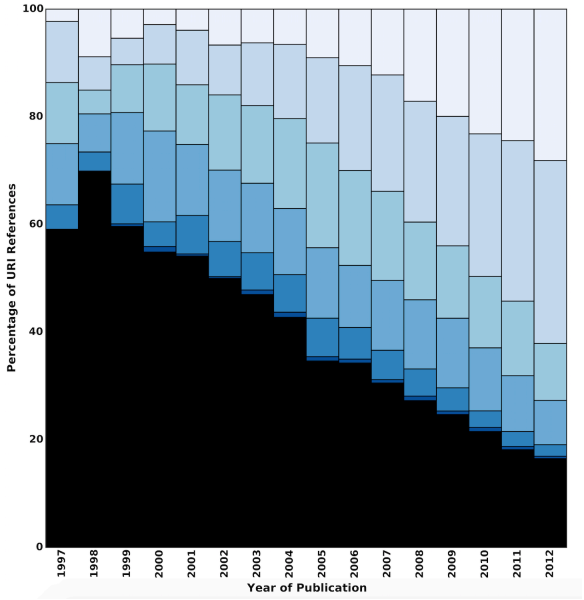
CC BY

Los Alamos NATIONAL LABORATORY

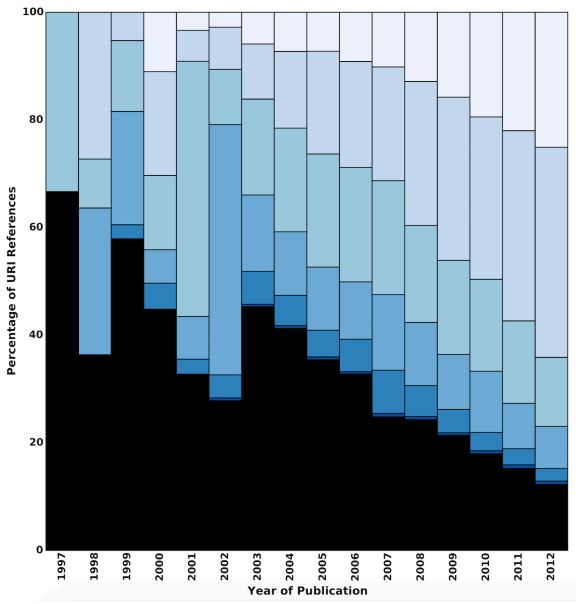
arXiv



Elsevier



PMC



CC BY

Los Alamos NATIONAL LABORATORY

Agenda

1. Definition: Reference Rot
2. Study: Quantification of Content Drift
3. Proposed Solution: Robust Links

CC BY

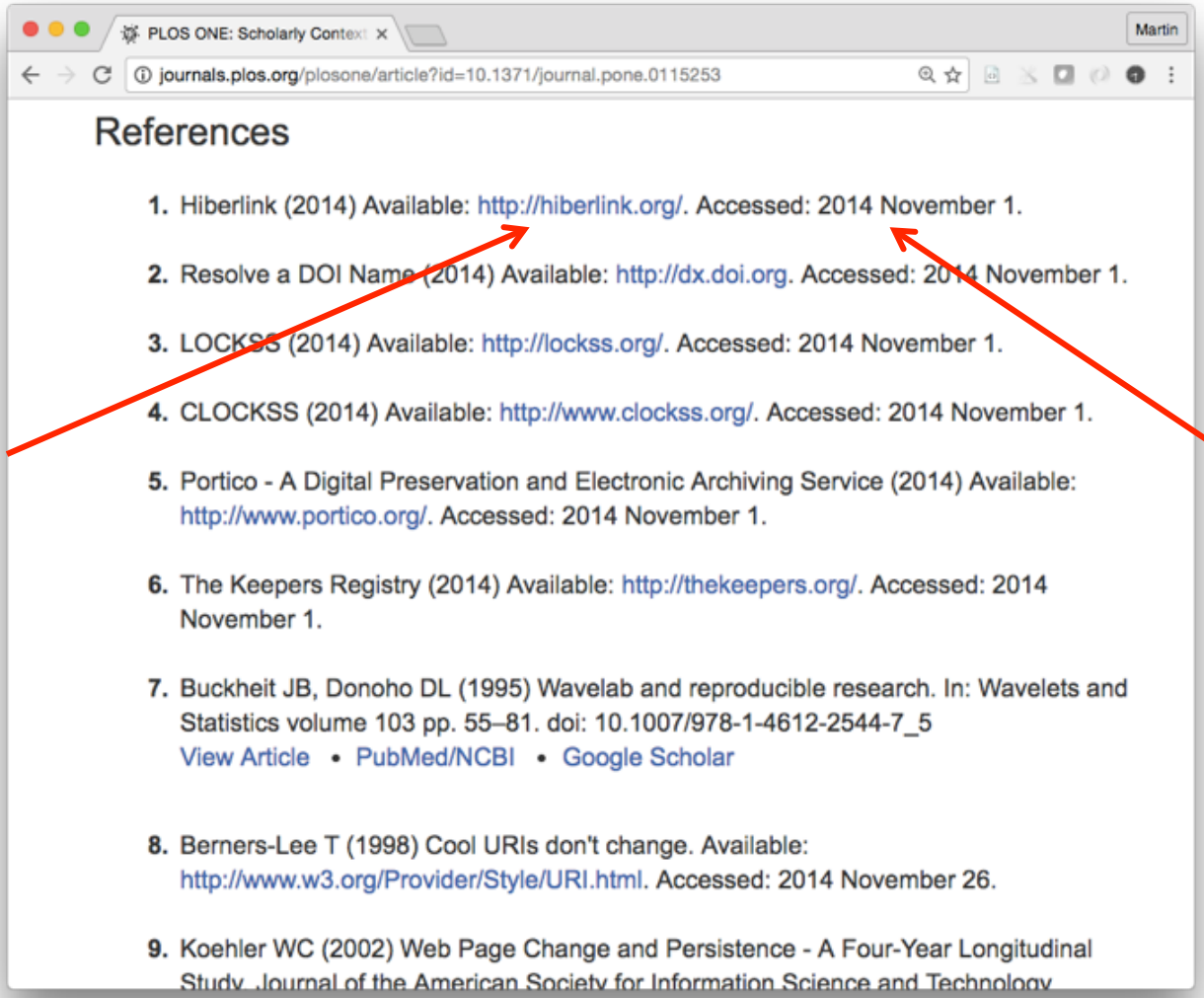
Why don't DOIs solve this problem?

- Designed to combat link rot of references to scholarly articles
- Fully relies on the custodians of DOI-identified resources to maintain links
- Strong incentives to invest in link stability, usability of their content is at stake
- Custodians of web at large resources do not have such incentives (typically web admins, not academic publishers)
- Not overly concerned about longevity of their website, the scholarly record
- Reference rot is a problem largely rooted outside scholarly communication community, will have to be solved by it

CC BY

Common Practice – original URI (& last accessed date)

Original URI



Accessed datetime

CC BY

Robust Links

1. **Create a snapshot** of referenced resources in a public web archive

perma.cc

archive.is
webpage capture

WebCite

CC BY

Common Practice – Archived URI

^ Heyerdahl 1961 However, Alfred Metraux pointed out that the rubble filled Rapanui walls were a fundamentally different design to those of the Inca, as these are trapezoidal in shape as opposed to the perfectly fitted rectangular stones of the Inca. See also this FAQ at the Wayback Machine (archived 11 October 2007)

URI of archived snapshot

Capture datetime

CC BY

Robust Links

1. Create a snapshot of referenced resources in a publically available web archive

2. **Decorate links** with:
   - URI of archived snapshot
   - datetime of archiving
   - resource's original URI

Benefits:

- Original URI allows finding captures in all web archives
- Capture datetime allows finding an appropriate capture in all web archives
- Uniform, machine-actionable

http://robustlinks.mementoweb.org/

CC BY

Robust Links - Desired Practice

URI of archived snapshot

^ Kirch, Patrick (2003). "Introduction to Pacific Islands Archaeology". Social Science Computing Laboratory, Berkeley. Archived from the original on 6 December 2008. Retrieved 21 December 2014.

Original URI

Capture datetime

CC BY

Link Decoration

```
<a href="https://www.cni.org/">CNI</a>
```

http://robustlinks.mementoweb.org/spec

CC BY

Link Decoration

```
<a href="https://www.cni.org/"

data-versionurl="http://archive.is/fItfw"
data-versiondate="2016-12-11">

CNI</a>
```

http://robustlinks.mementoweb.org/spec

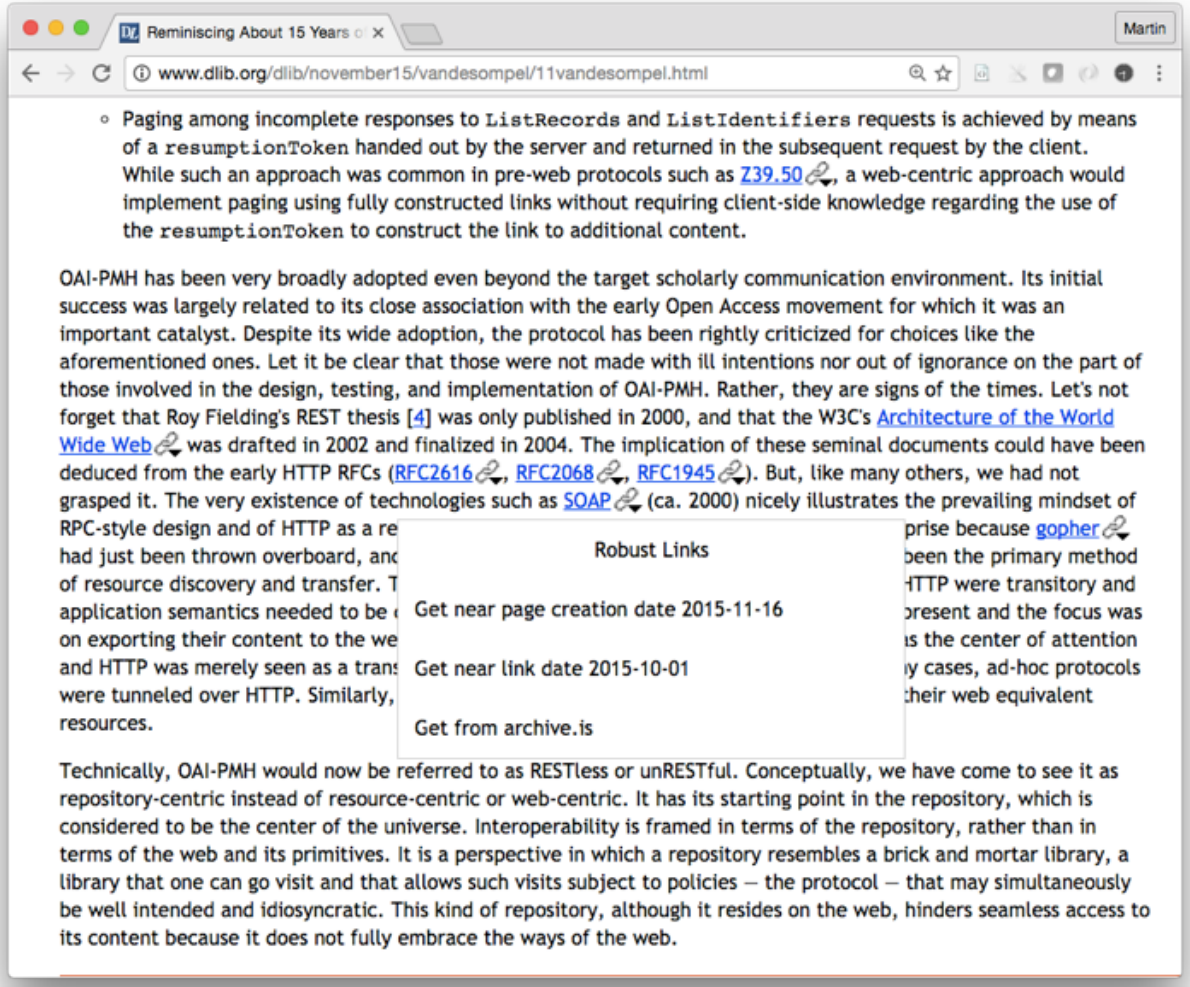
CC BY

Link Decoration in Action

http://dx.doi.org/10.1045/november2015-vandesompel

CC BY

Link Decoration in Action



http://dx.doi.org/10.1045/november2015-vandesompel

CC BY

Take-Aways

1. Scholarly articles increasingly contain URI references to web at large resources.

2. Such resources are subject to reference rot (link rot + content drift).

3. Custodians of these resources are typically not overly concerned with archiving of their content and longevity of the scholarly record.

4. Authors, publishers, and other parties can help tackle this problem by making links more robust.

CC BY

Reference Rot in Scholarly Communication: A Reliable Quantification and a Proposed Solution

Martin Klein
@mart1nkle1n

Research Library
Los Alamos National Laboratory

CC BY

Los Alamos NATIONAL LABORATORY