Investigating Reference Rot in Web-Based Scholarly Communication

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp

http://hiberlink.org #hiberlink
http://mementoweb.org #memento

Hiberlink is funded by the Andrew W. Mellon Foundation

Hiberlink Team Work

- Los Alamos National Laboratory:
  - Research Library: Martin Klein, Harihar Shankar, **Herbert Van de Sompel**

- University of Edinburgh:
  - Edina: **Peter Burnhill**, Neil Mayo, Muriel Mewissen, Christine Rees, Tim Stickland, Riachard Wincewicz
  - Language Technology Group: Beatrice Alex, **Claire Grover**, Richard Tobin, Ke “Adam” Zhou

- Funding: Andrew W. Mellon Foundation

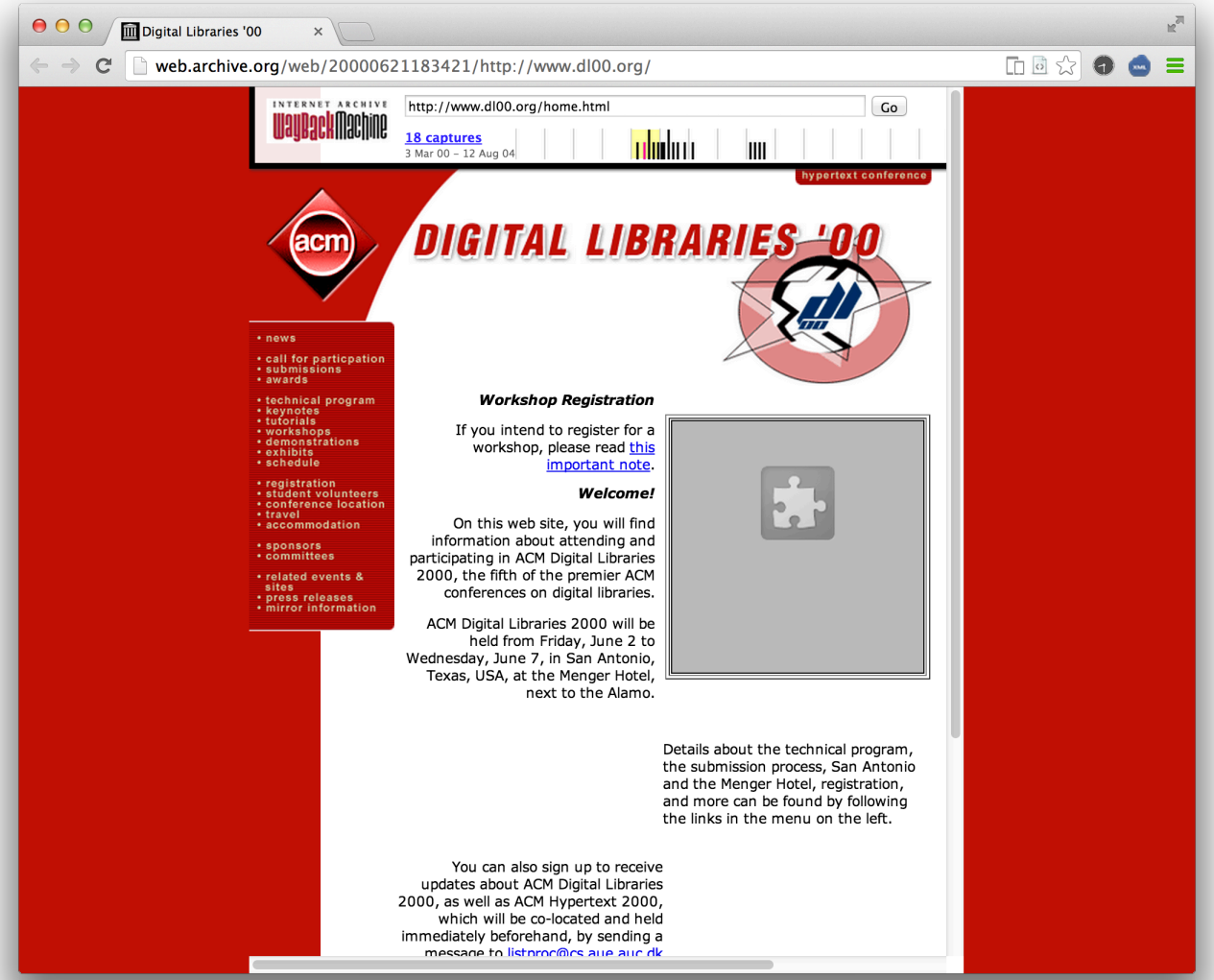
Reference Rot

Link Rot


IKEA
The page cannot be displayed
The page you are looking for is currently unavailable. The site might be experiencing technical difficulties, or you ma
to adjust your browser settings.


IKEA
The page cannot be displayed
The page you are looking for is currently unavailable. The site might be experiencing technical difficulties, or you ma
to adjust your browser settings.

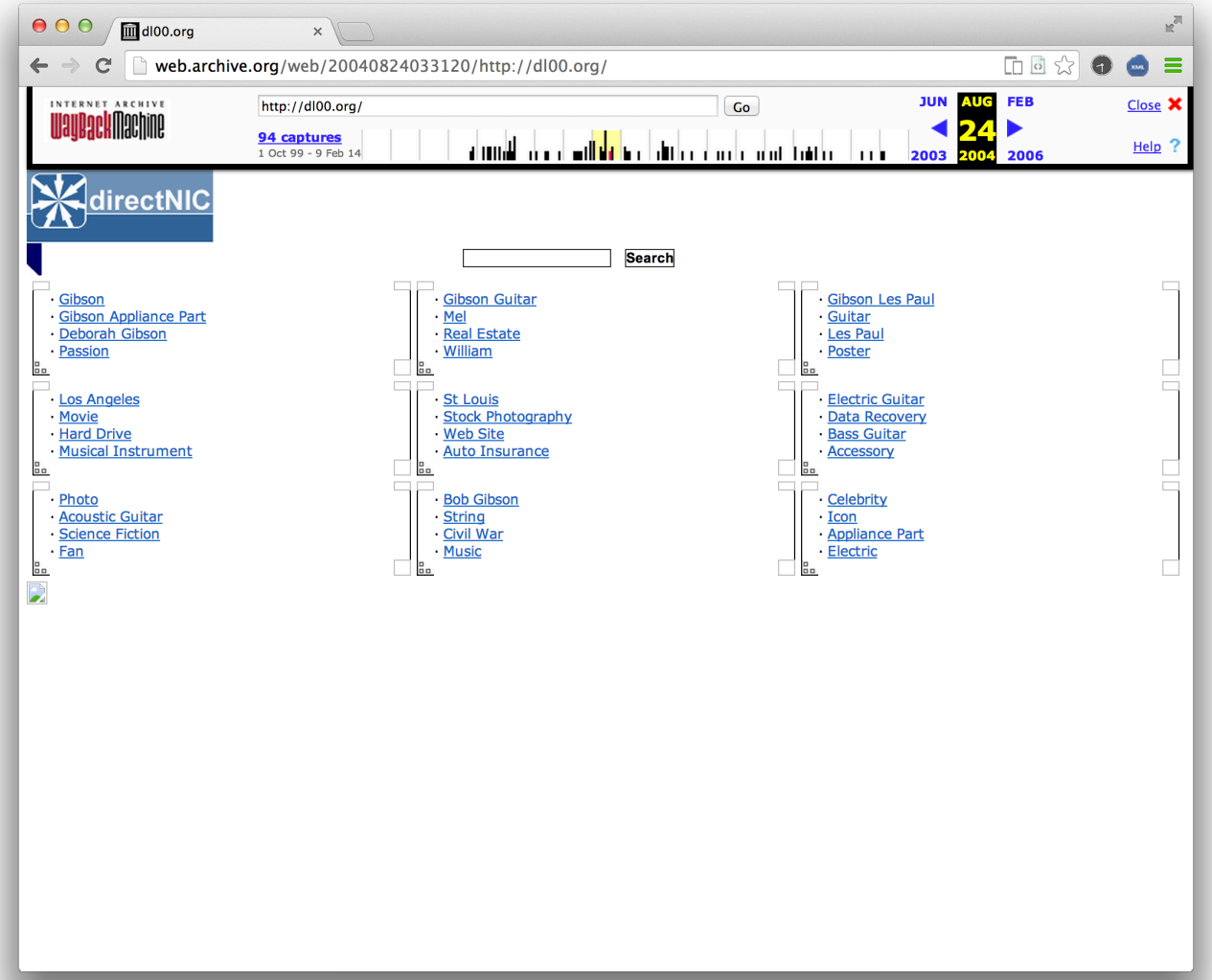
Content Drift



http://dl00.org

2000

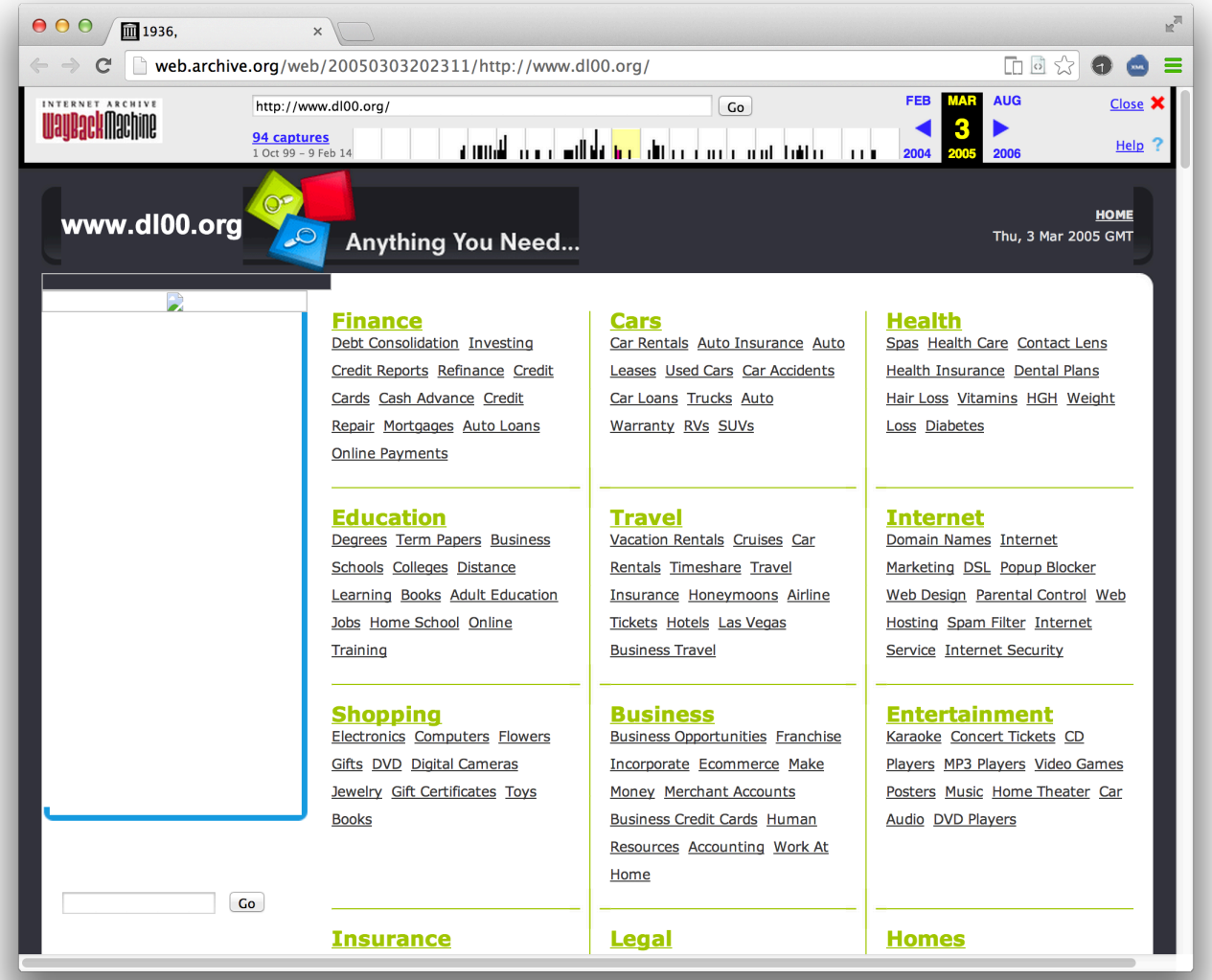
Content Drift



http://dl00.org

2004

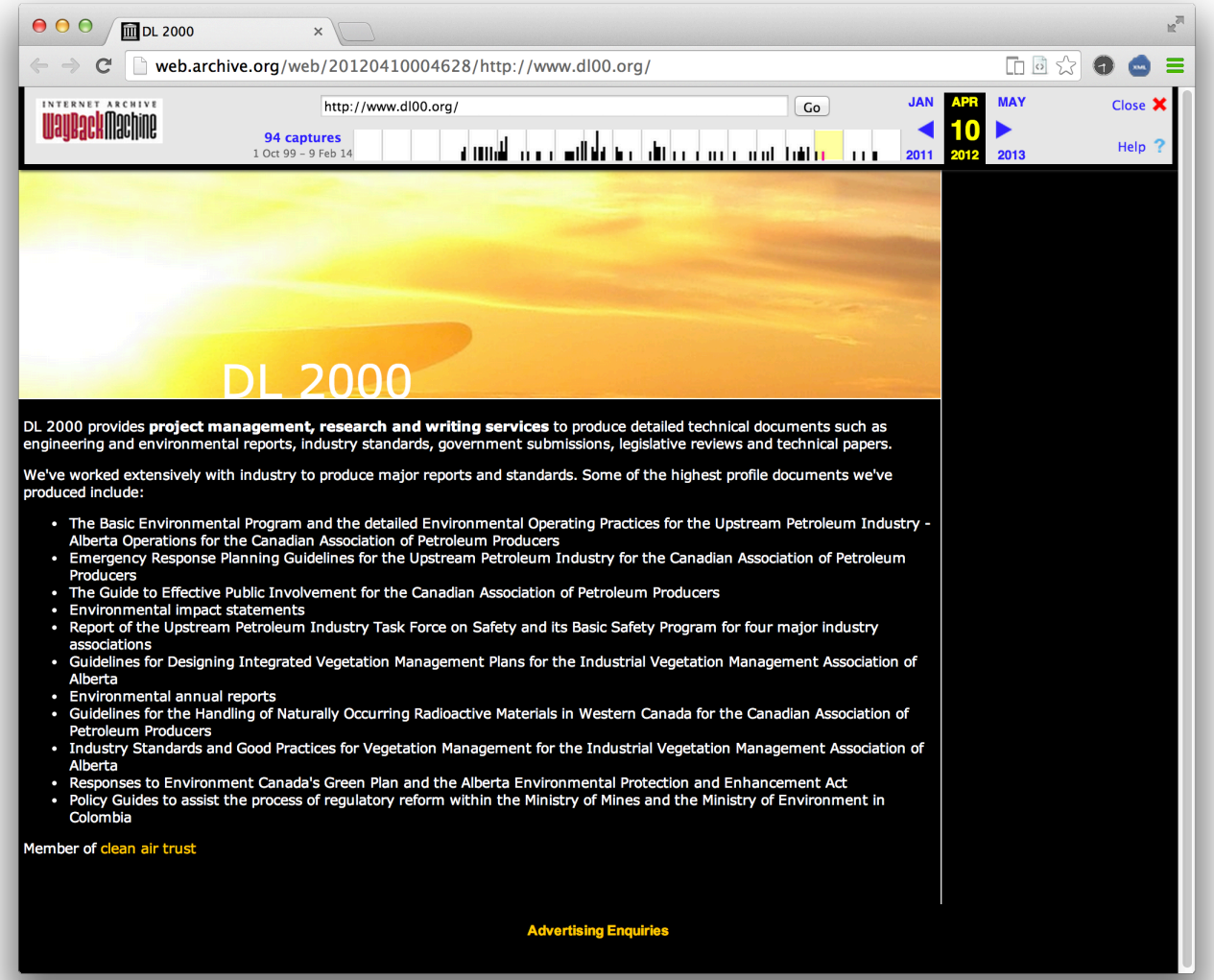
Content Drift



http://dl00.org

2005

Content Drift



http://dl00.org

2008

The New York Times Cares

SIDEBAR

In Supreme Court Opinions, Web Links to Nowhere

By ADAM LIPTAK
Published: September 23, 2013

WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to a new study, 49 percent of the hyperlinks in Supreme Court decisions no longer work.

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT
REPRINTS

Enlarge This Image

Stephan Savoia/Associated Press
Justice Samuel A. Alito Jr.

This can sometimes be amusing. A link in a 2011 Supreme Court opinion about violent video games by Justice Samuel A. Alito Jr. now leads to a mischievous error message.

"Aren't you glad you didn't cite to this Web page?" it asks. "If you had, like Justice Alito did, the original content would have long since 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."

The prankster has a point. The modern Supreme Court opinion is increasingly built on sand.

Hyperlinks are a huge and welcome convenience, of course, said Jonathan Zittrain, who teaches law and computer science at Harvard and who prepared the study with Kendra Albert, a law student there. "Things are readily accessible," he said, "until they aren't."

Links in Supreme Court decisions:

- Link rot: **29%**
- Reference rot: **49.9%**

http://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html

Entrance Hiberlink

- These resources:
  - Are not necessarily under the custodianship of parties that care about long term integrity, access
  - Do not necessarily have the same sense of fixity that e.g. journal articles have

- Links to these resources are subject to Reference Rot:
  - Link Rot: Link stops working, e.g. HTTP 404
  - Content Drift: Linked content changes over time

D-Lib Magazine
September 2004

Volume 10 Number 9

ISSN 1082-9873

Rethinking Scholarly Communication

Building the System that Scholars Deserve

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
<herbertv@lanl.gov>

Sandy Payette
Cornell University, Computing and Information Science
<payette@cs.cornell.edu>

John Erickson
Hewlett-Packard Laboratories, Digital Media Systems Lab
<john.erickson@hp.com>

Carl Lagoze
Cornell University, Computing and Information Science
<lagoze@cscornell.edu>

Simeon Warner
Cornell University, Computing and Information Science
<simeon@cs.cornell.edu>

References

!Exist	Atkins, D. et al.. 2003. National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, *Revolutionizing Science and Engineering through Cyber-infrastructure*, <http://www.communitytechnology.org/nsf_ci_report/>.	Archived
Exist	Brody, T., Kampa, S., Harnad, S., Carr, L. and Hitchcock, S. 2003. Digitometric Services for Open Archives Environments. In *Proceedings of European Conference on Digital Libraries 2003*, pages pp. 207-220, Trondheim, Norway. <http://eprints.ecs.soton.ac.uk/archive/00007503/>.	Archived

Frey, J., De Roure, D. and Carr, L. 2002. *Publication at Source: Scientific Communication from a Publication Web to a Data Grid*. <http://eprints.ecs.soton.ac.uk/archive/00007852/>.

Henry, G. 2003. On-line publishing in the 21-st Century: Challenges and Opportunities. *D-Lib Magazine*, Volume 9, Issue 10. <doi:10.1045/october2003-henry>.

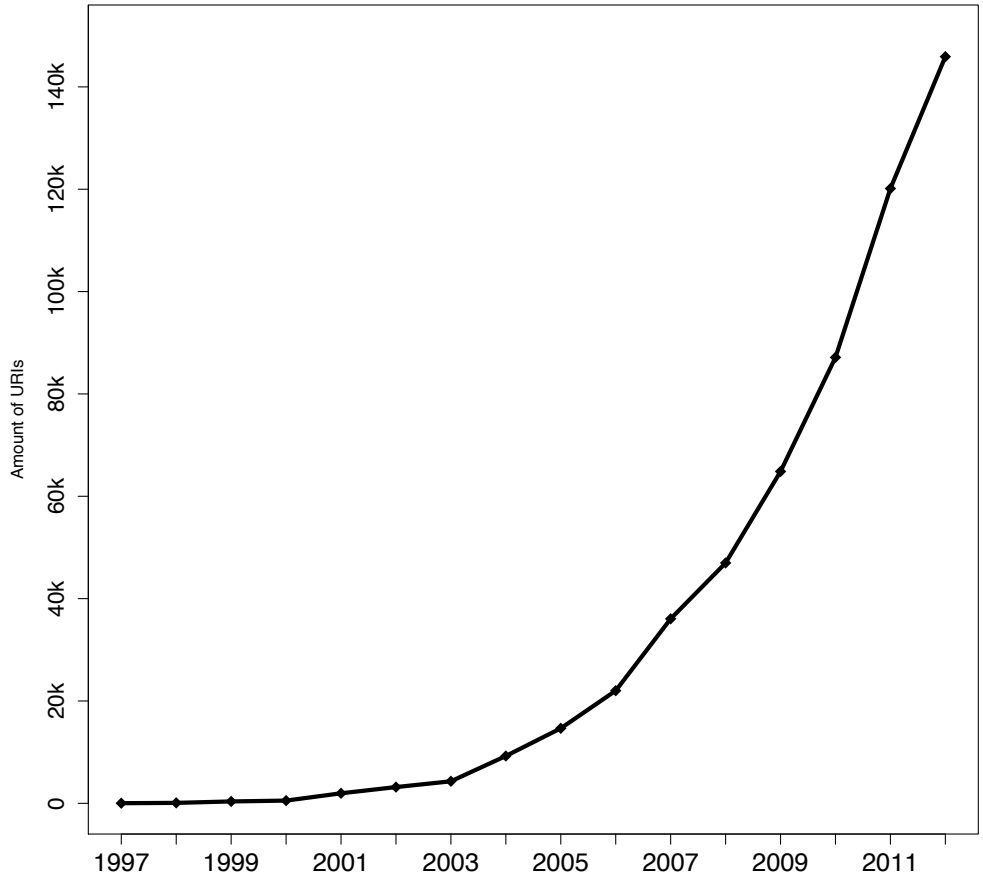
!Exist	Lynch, C. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report* 226. February 2003, <http://www.arl.org/newsltr/226/ir.html>.	Archived
!Exist	Payette, S., and Staples, T. 2002. The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services. *European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, September 2002. <http://www.fedora.info/documents/ecdl2002final.pdf>.	!Archived

Pöschl, U. 2004. Interactive Journal Concept for Improved Scientific Publishing and Quality Assurance. *Learned Information*, Volume 17, Number 2, pp 105-113. <doi:10.1087/095315104322958481>.

Reich, V. and Rosenthal, D. 2001. LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*, Volume 7, Issue 6. <doi:10.1045/june2001-reich>.

Exist	Roosendaal, H., and Geurts, P. 1997. Forces and functions in scientific communication: an analysis of their interplay. *Cooperative Research Information Systems in Physics*, August 31—September 4 1997, Oldenburg, Germany. <http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>.	Archived

Articles Increasingly Link to Web Resources



URIs extracted from PubMed Central papers

hiberlink

Quantifying Reference Rot

Hiberlink - Martin Klein
IIPC GA, Paris, France, May 19th 2014

Los Alamos NATIONAL LABORATORY

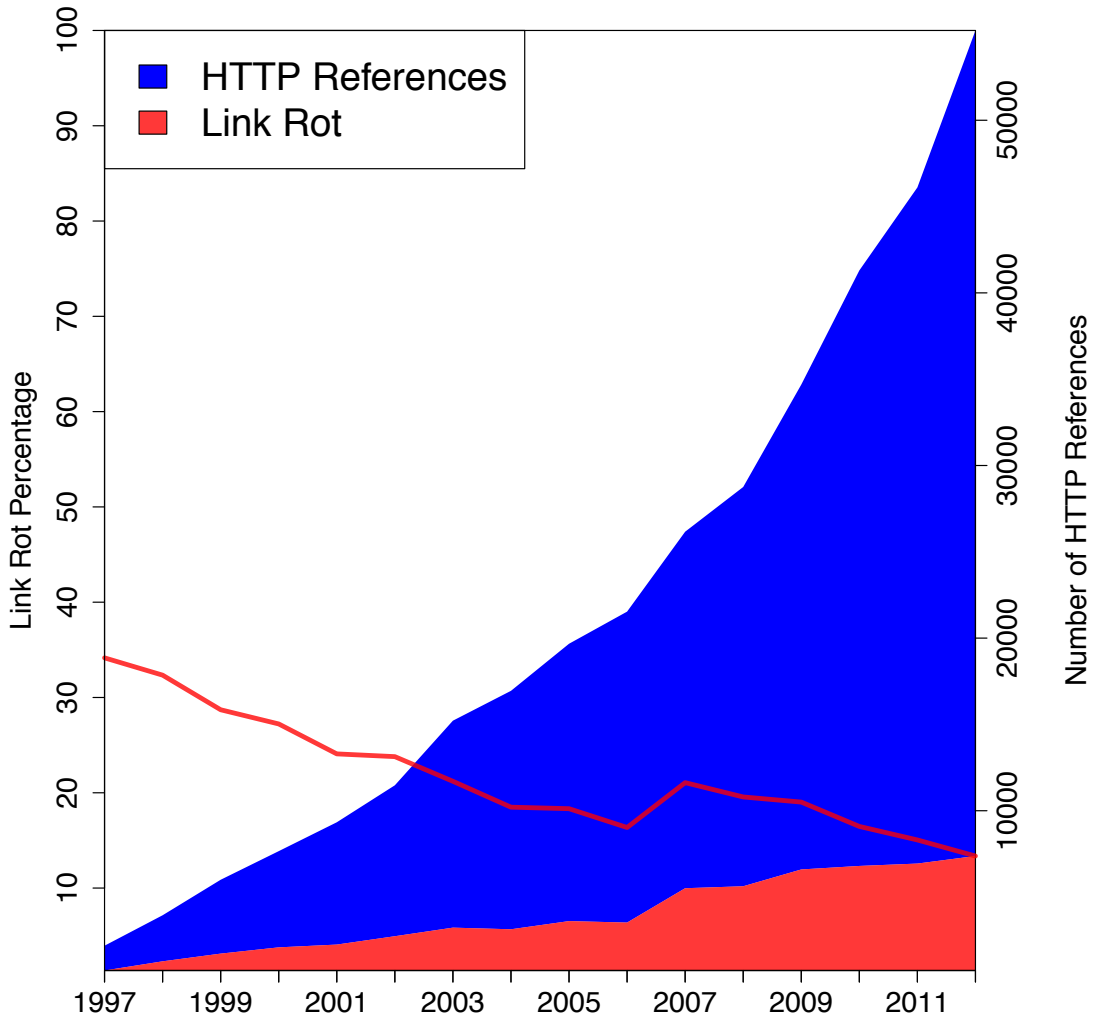
THE UNIVERSITY of EDINBURGH

Study Parameters

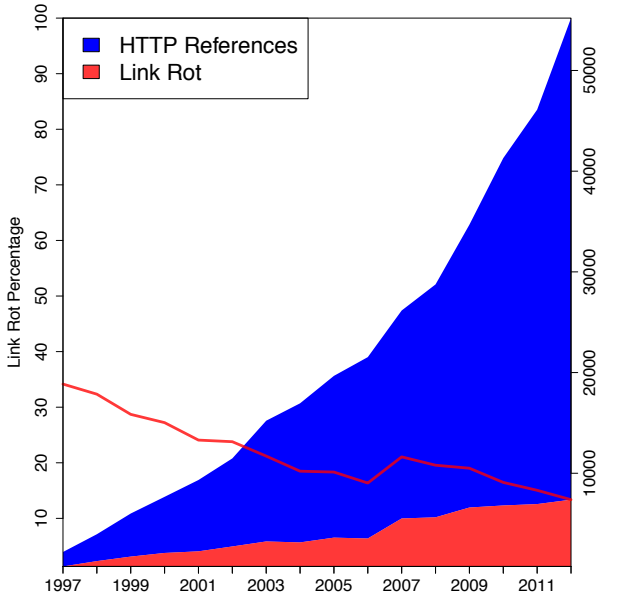
- Time frame of publications: Jan 1997 – Dec 2012
- Articles in XML and PDF format
  - Convert PDF to XML
  - URI extraction
    - Challenge: URI broken up by newline; underscore as image
  - Store publication date
- URI live web test
- URI archive lookup via Memento infrastructure

	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

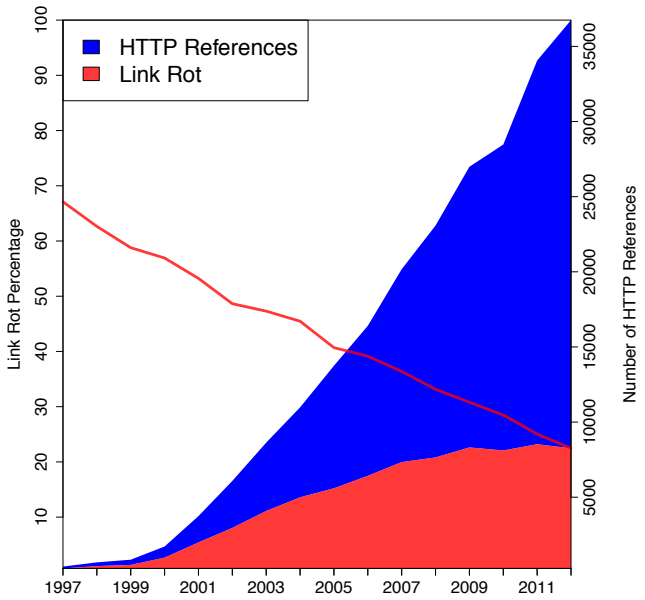
Link Rot in arXiv



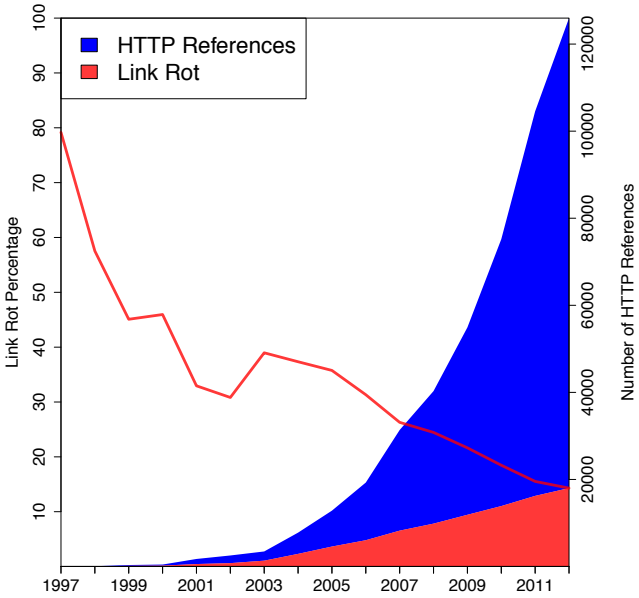
arXiv



Elsevier

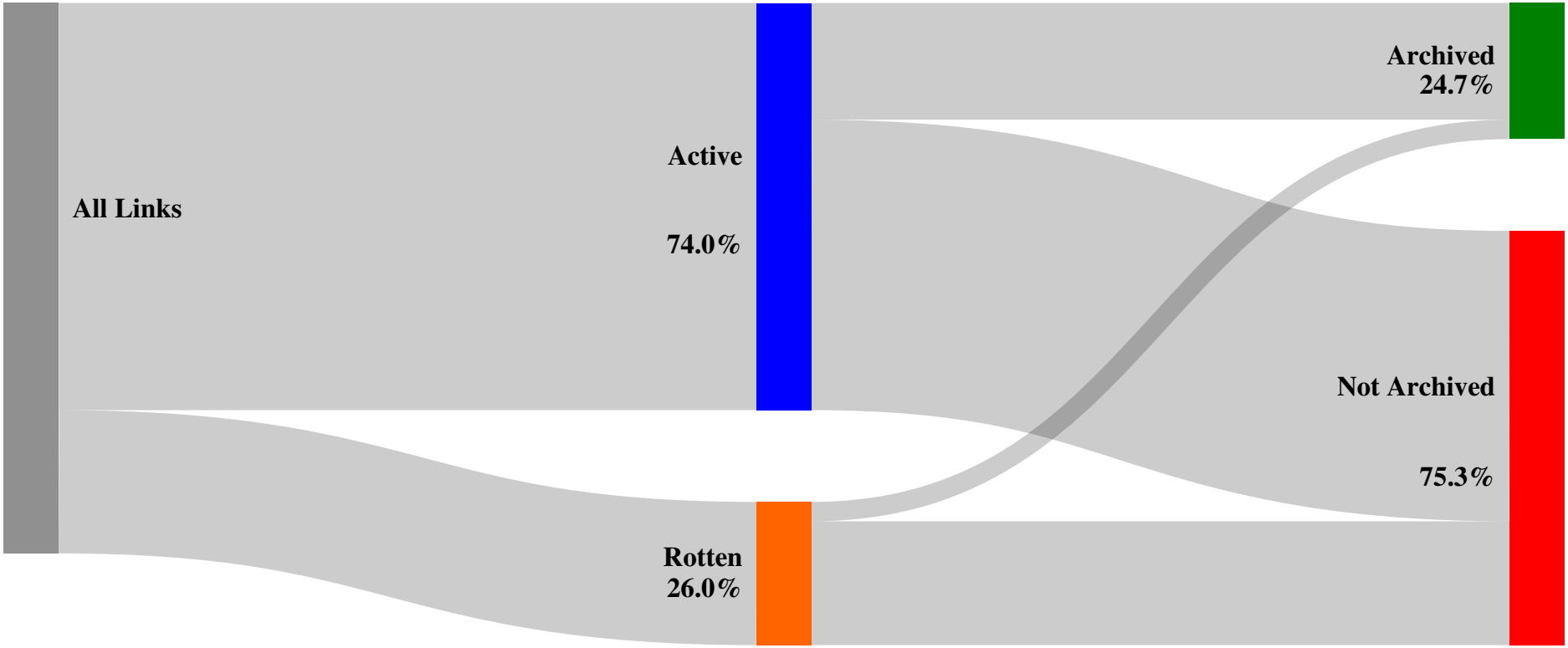


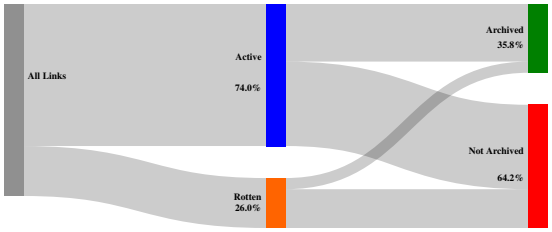
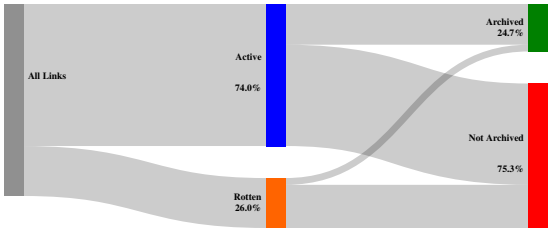
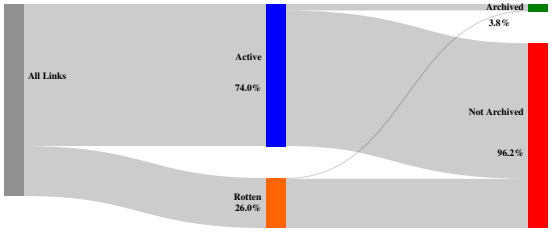
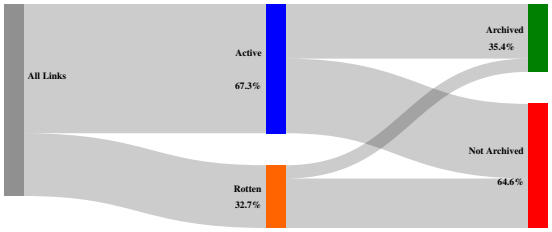
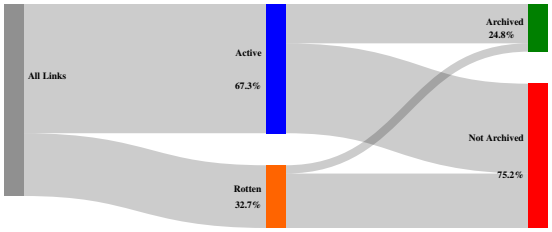
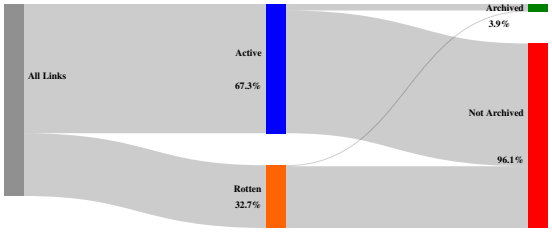
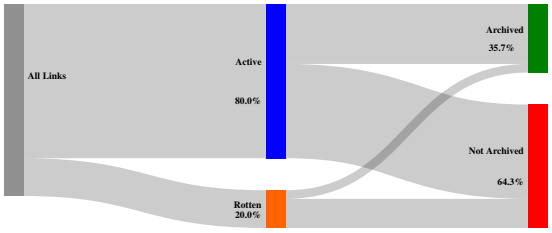
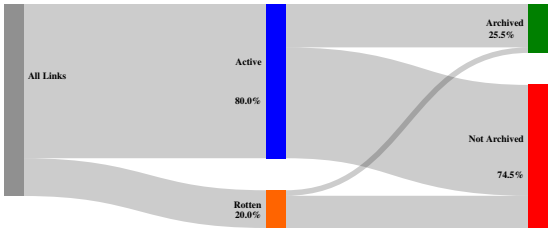
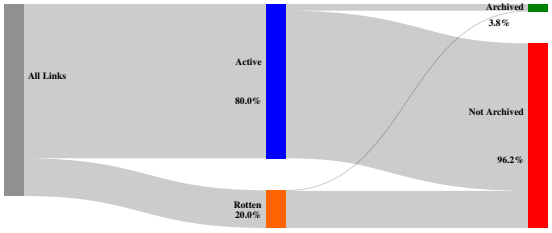
PMC



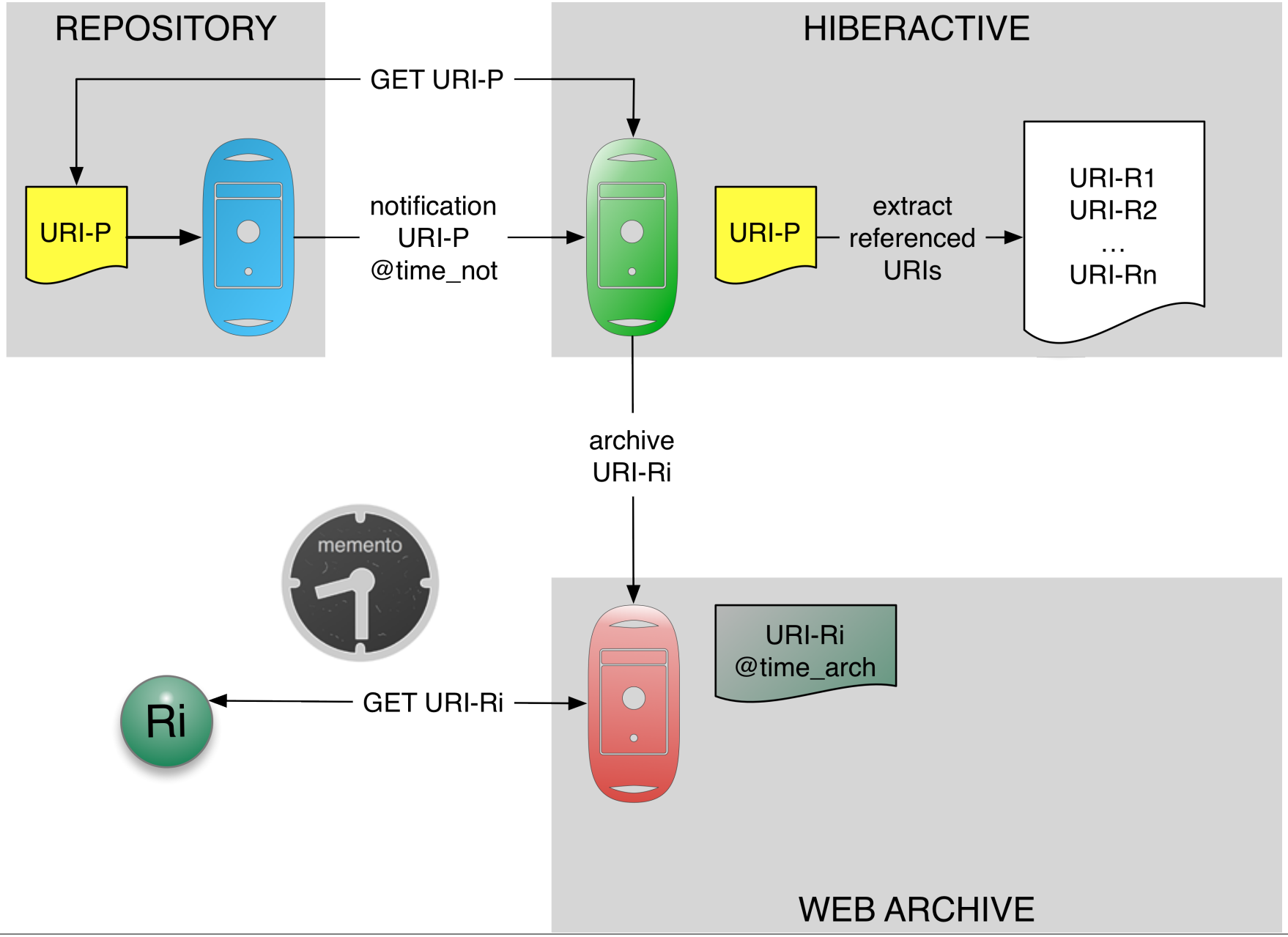
Content Drift in arXiv

Archived within 14 days of publication



	1 Month	14 Days	24 Hours
arXiv			
Elsevier			
PMC			

Solving Reference Rot



Linking to Archived Resources

- Link by means of the original URI
- Augment the link with temporal context aimed at increasing link robustness
  - Date of linking
  - URI of archived snapshot(s)
- **404-No-More** collaboration aims at standardizing an approach for HTML
  - Harvard Law Library (perma.cc)
  - Harvard Berkman Center for Internet & Security
  - Los Alamos National Laboratory
  - Old Dominion University

HTML Links

```
<a href="http://www.bnf.fr">
   Link to the BNF
</a>
```

mset – Augmenting Links

```
<a href="http://www.bnf.fr"
   mset="2014-05-19,
   http://archive.today/zdpAn">
   Link to the BNF
</a>
```

mset – Augmenting Links

```
<a href="http://www.bnf.fr"
   mset="2014-05-19,
   http://archive.today/zdpAn 2014-05-15 memento">
   Link to the BNF
</a>
```

mset – Augmenting Links

```
<a href="http://www.bnf.fr"
   mset="2014-05-19,
   http://archive.today/zdpAn,
   http://perma.cc/SC89-PAHK">
   Link to the BNF
</a>
```

Investigating Reference Rot in Web-Based Scholarly Communication

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp

http://hiberlink.org #hiberlink
http://mementoweb.org #memento

Hiberlink is funded by the Andrew W. Mellon Foundation