



Investigating Reference Rot in Web-Based Scholarly Communication

**Martin Klein**

Los Alamos National Laboratory  
@mart1nkle1n

Herbert Van de Sompel

Los Alamos National Laboratory  
@hvdsomp

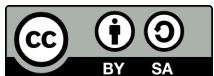
<http://hiberlink.org> #hiberlink  
<http://mementoweb.org> #memento

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





# Hiberlink

## Reference Rot



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



 **Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# Link Rot



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



Los Alamos  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH



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



**Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# Content Drift

<http://dl00.org>  
2000

INTERNET ARCHIVE  
Wayback Machine  
18 captures  
3 Mar 00 - 12 Aug 04

http://www.dl00.org/home.html

hypertext conference

acm

**DIGITAL LIBRARIES '00**

- news
- call for participation
- submissions
- awards
- technical program
- keynotes
- tutorials
- workshops
- demonstrations
- exhibits
- schedule
- registration
- student volunteers
- conference location
- travel
- accommodation
- sponsors
- committees
- related events & sites
- press releases
- mirror information

**Workshop Registration**

If you intend to register for a workshop, please read [this important note](#).

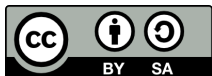
**Welcome!**

On this web site, you will find information about attending and participating in ACM Digital Libraries 2000, the fifth of the premier ACM conferences on digital libraries.

ACM Digital Libraries 2000 will be held from Friday, June 2 to Wednesday, June 7, in San Antonio, Texas, USA, at the Menger Hotel, next to the Alamo.

Details about the technical program, the submission process, San Antonio and the Menger Hotel, registration, and more can be found by following the links in the menu on the left.

You can also sign up to receive updates about ACM Digital Libraries 2000, as well as ACM Hypertext 2000, which will be co-located and held immediately beforehand, by sending a message to [listserv@cs.auc.dk](mailto:listserv@cs.auc.dk)



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



Los Alamos  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# Content Drift

<http://dl00.org>  
2004

Internet Archive Wayback Machine  
94 captures  
1 Oct 99 - 9 Feb 14

http://dl00.org/

directNIC

Search

- Gibson
  - Gibson Appliance Part
  - Deborah Gibson
  - Passion
- Los Angeles
  - Movie
  - Hard Drive
  - Musical Instrument
- Photo
  - Acoustic Guitar
  - Science Fiction
  - Fan
- Gibson Guitar
  - Mei
  - Real Estate
  - William
- St Louis
  - Stock Photography
  - Web Site
  - Auto Insurance
- Bob Gibson
  - String
  - Civil War
  - Music
- Gibson Les Paul
  - Guitar
  - Les Paul
  - Poster
- Electric Guitar
  - Data Recovery
  - Bass Guitar
  - Accessory
- Celebrity
  - Icon
  - Appliance Part
  - Electric

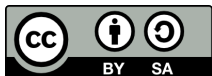


# Content Drift

http://dl00.org  
2005

The screenshot shows a browser window with the URL <http://www.dl00.org/>. The page is titled "www.dl00.org Anything You Need..." and is dated "Thu, 3 Mar 2005 GMT". The main content is organized into several columns, each with a category heading and a list of sub-links:

- Finance**: Debt Consolidation, Investing, Credit Reports, Refinance, Credit Cards, Cash Advance, Credit Repair, Mortgages, Auto Loans, Online Payments
- Education**: Degrees, Term Papers, Business Schools, Colleges, Distance Learning, Books, Adult Education, Jobs, Home School, Online Training
- Shopping**: Electronics, Computers, Flowers, Gifts, DVD, Digital Cameras, Jewelry, Gift Certificates, Toys, Books
- Insurance**
- Cars**: Car Rentals, Auto Insurance, Auto Leases, Used Cars, Car Accidents, Car Loans, Trucks, Auto Warranty, RVs, SUVs
- Travel**: Vacation Rentals, Cruises, Car Rentals, Timeshare, Travel Insurance, Honeymoons, Airline Tickets, Hotels, Las Vegas, Business Travel
- Business**: Business Opportunities, Franchise, Incorporate, Ecommerce, Make Money, Merchant Accounts, Business Credit Cards, Human Resources, Accounting, Work At Home
- Legal**
- Health**: Spas, Health Care, Contact Lens, Health Insurance, Dental Plans, Hair Loss, Vitamins, HGH, Weight Loss, Diabetes
- Internet**: Domain Names, Internet Marketing, DSL, Popup Blocker, Web Design, Parental Control, Web Hosting, Spam Filter, Internet Service, Internet Security
- Entertainment**: Karaoke, Concert Tickets, CD Players, MP3 Players, Video Games, Posters, Music, Home Theater, Car Audio, DVD Players
- Homes**



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

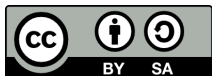
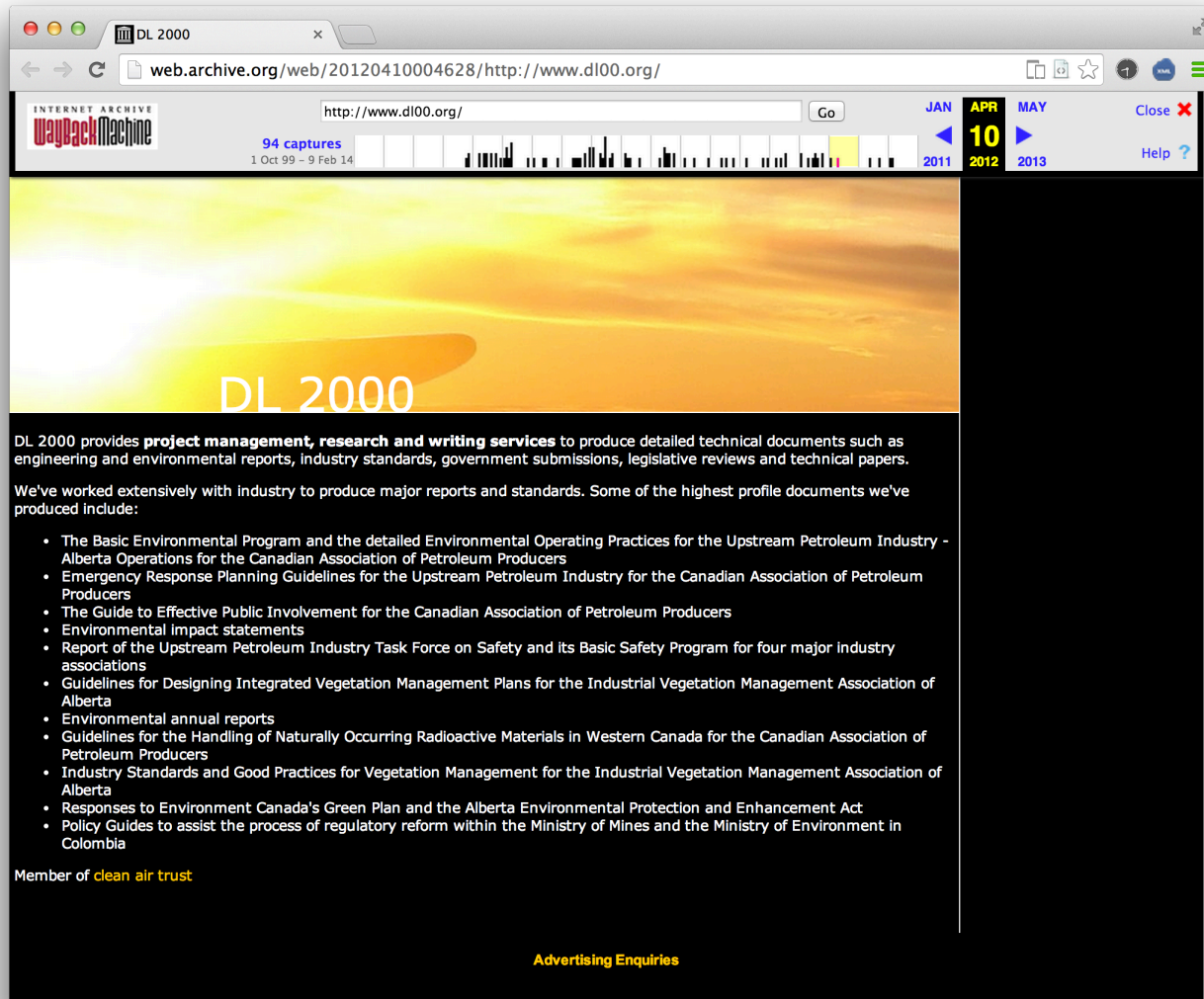


Los Alamos  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH



# Content Drift

<http://dl00.org>  
2008



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



**Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# 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.



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

[Enlarge This Image](#)

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.”

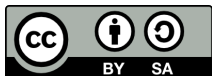
- FACEBOOK
- TWITTER
- GOOGLE+
- SAVE
- E-MAIL
- SHARE
- PRINT
- REPRINTS

BLACK NATIVITY  
NOVEMBER 27  
WATCH TRAILER

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>



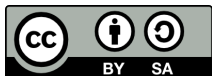
Hiberlink - Martin Klein  
IIPC GA, Paris, France, May 19th 2014



Los Alamos  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# 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](#)

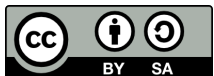
Hewlett-Packard Laboratories, Digital Media Systems Lab  
<john.erickson@hp.com>

[Carl Lagoze](#)

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

[Simeon Warner](#)

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



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



  
**Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

## 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/](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](https://doi.org/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**

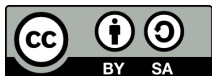
Pöschl, U. 2004. Interactive Journal Concept for Improved Scientific Publishing and Quality Assurance. *Learned Information*, Volume 17, Number 2, pp 105-113. <[doi:10.1087/095315104322958481](https://doi.org/10.1087/095315104322958481)>.

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](https://doi.org/10.1045/june2001-reich)>.

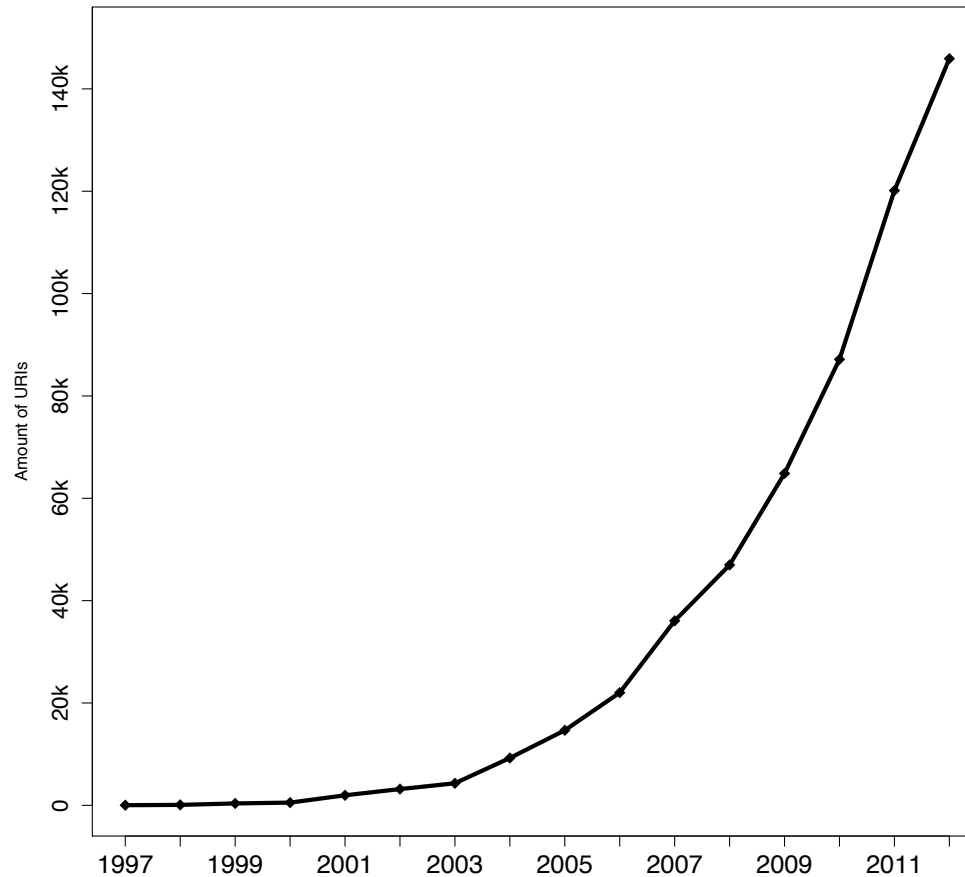
**Exist**

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

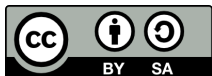


 **Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH

# Study Parameters

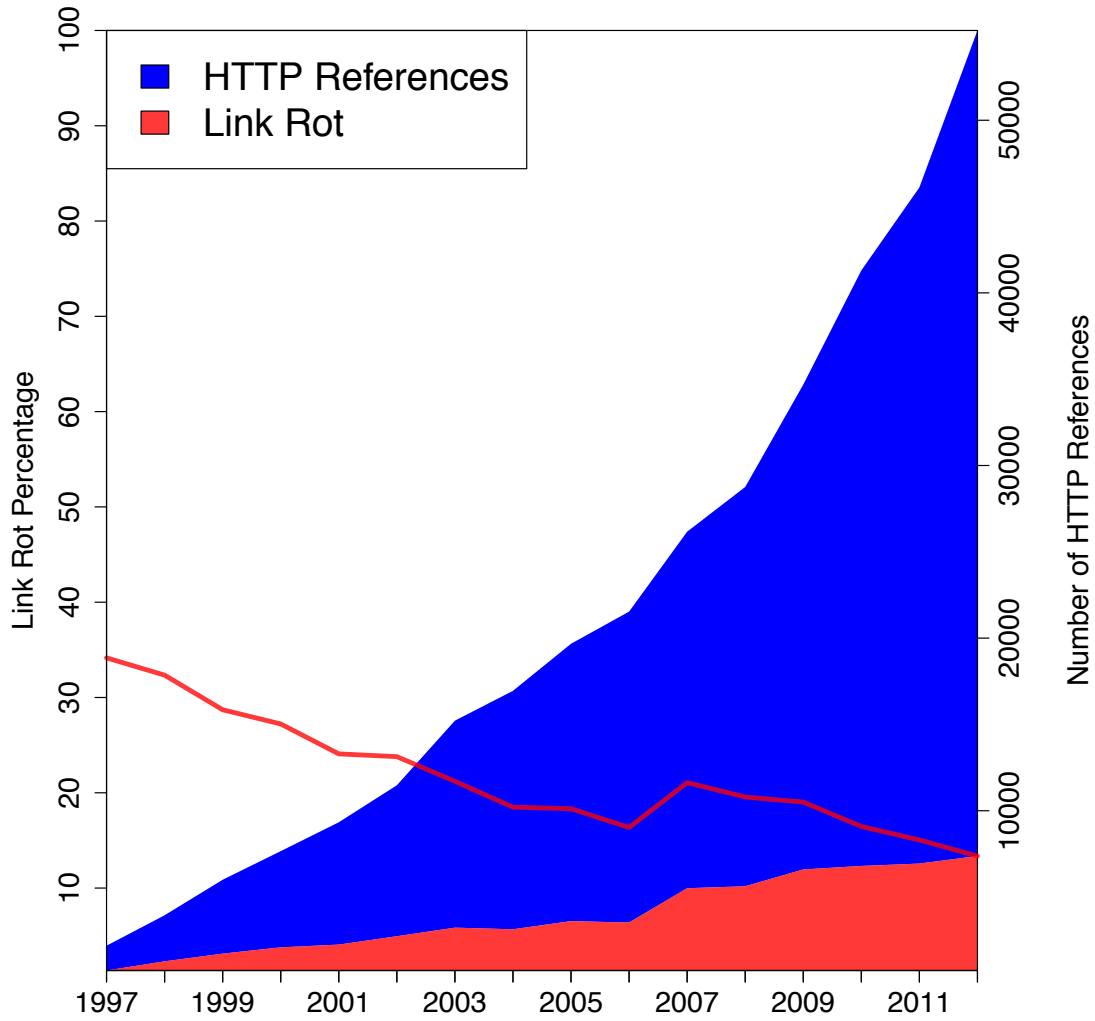
- Time frame of publications: Jan 1997 – Dec 2012
- Articles in XML and PDF format
  - Convert PDF to XML
  - URI extraction
    - Challenge: URI broken up by newline; underscore as image
  - Store publication date
- URI live web test
- URI archive lookup via Memento infrastructure

	<b>arXiv</b>	<b>Elsevier</b>	<b>PMC</b>
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





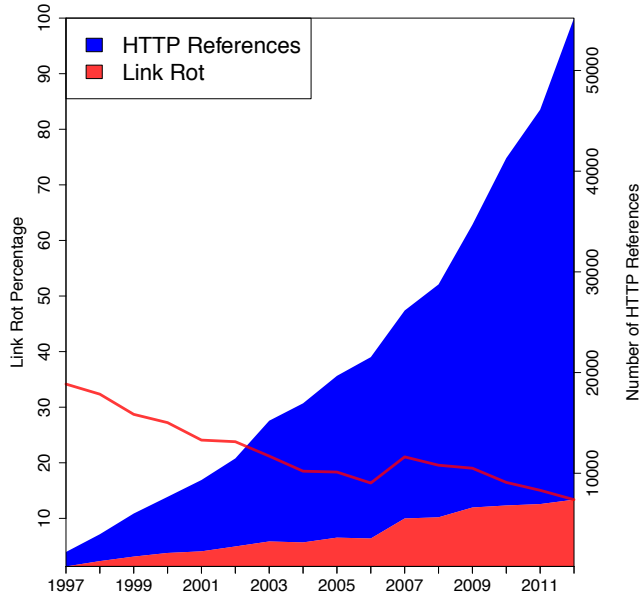
# Link Rot in arXiv



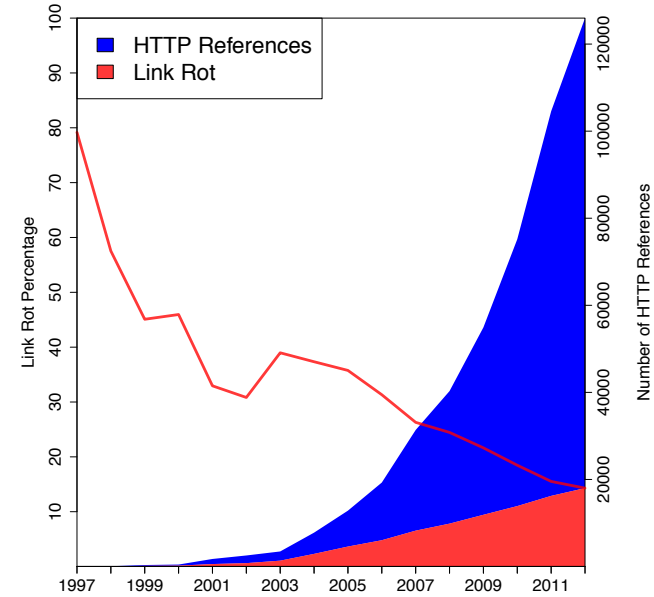
Hiberlink - Martin Klein  
IIPC GA, Paris, France, May 19th 2014



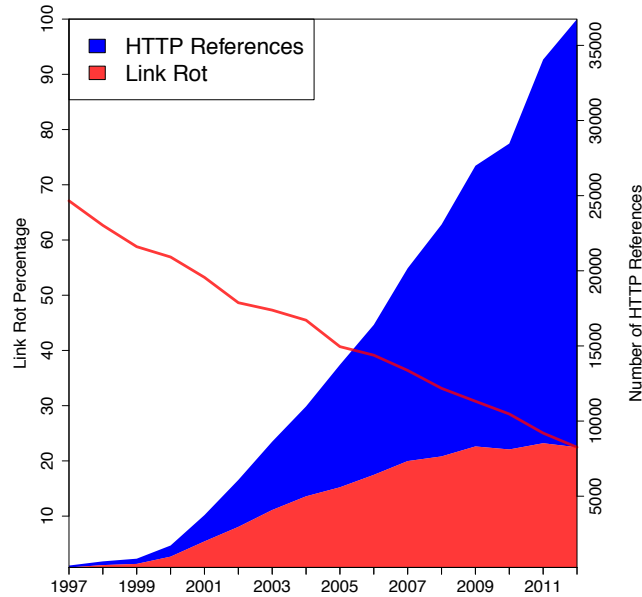
# arXiv



# PMC

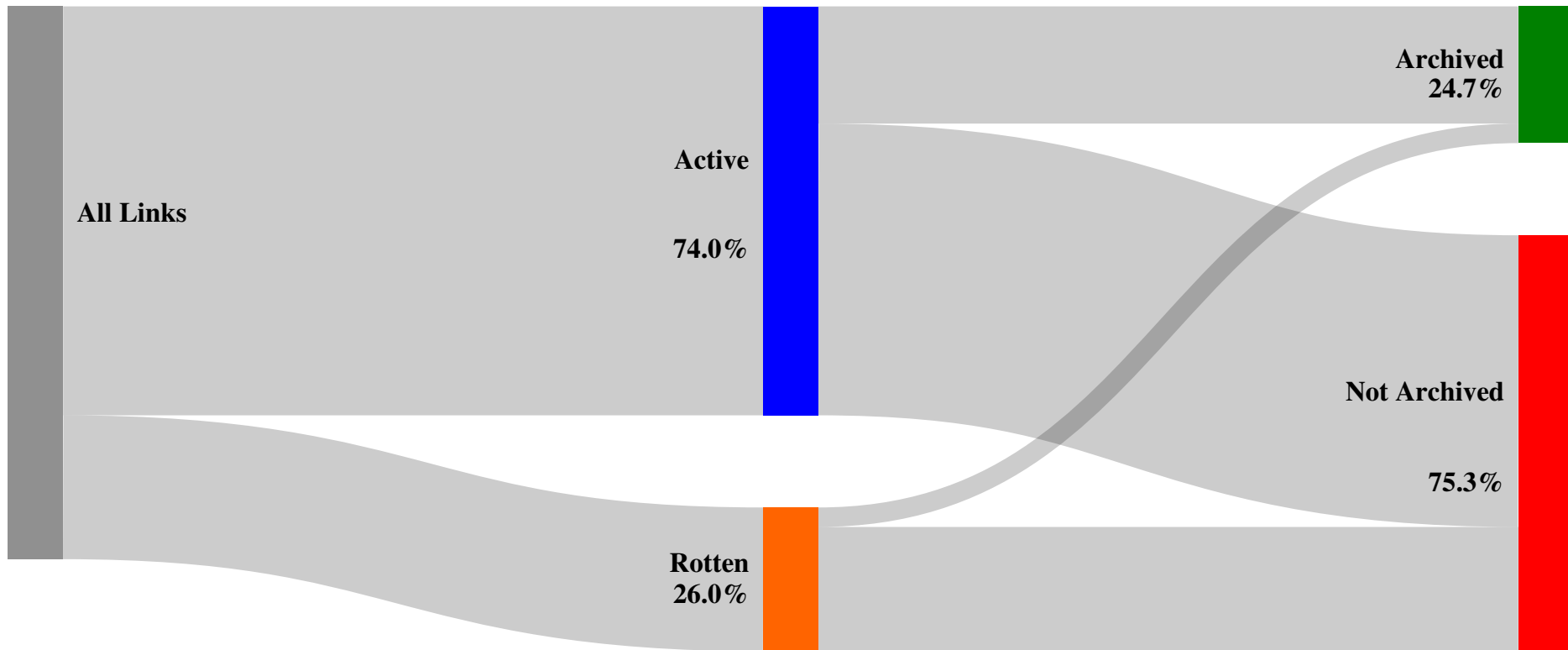


# Elsevier



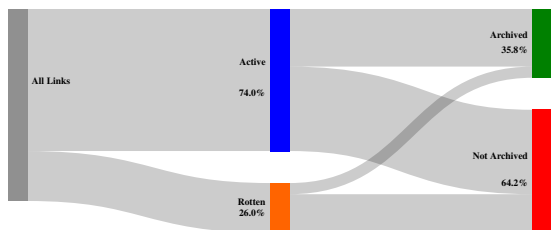
# Content Drift in arXiv

## Archived within 14 days of publication

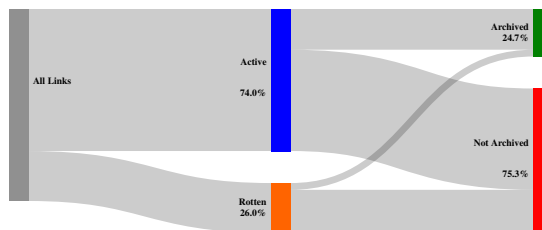


arXiv

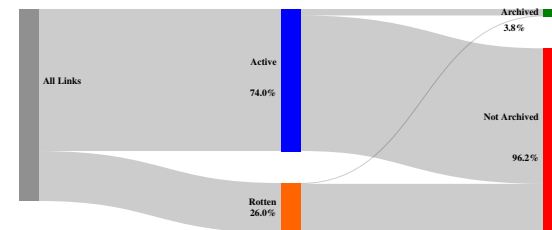
1 Month



14 Days

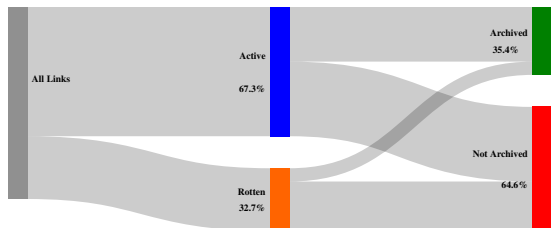


24 Hours

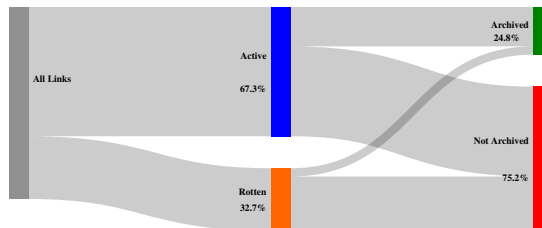


Elsevier

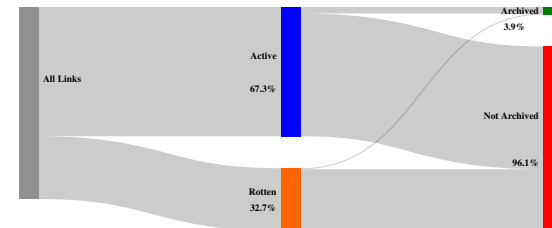
1 Month



14 Days

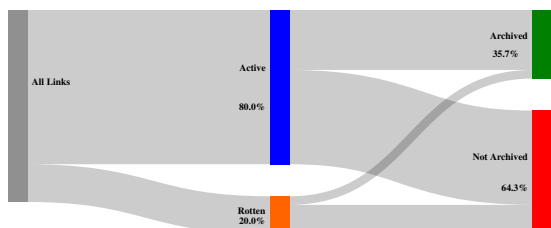


24 Hours

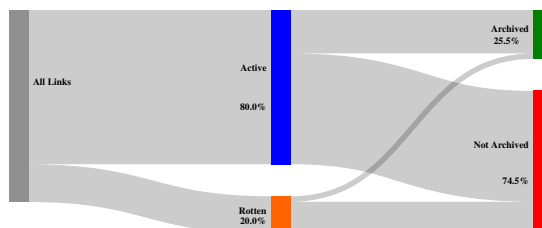


PMC

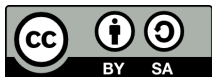
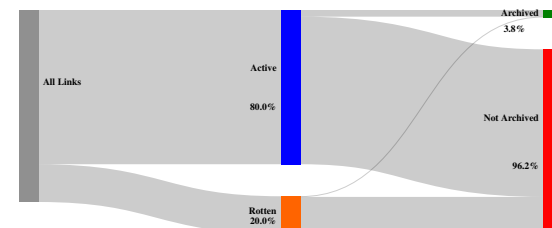
1 Month



14 Days



24 Hours





# Hiberlink

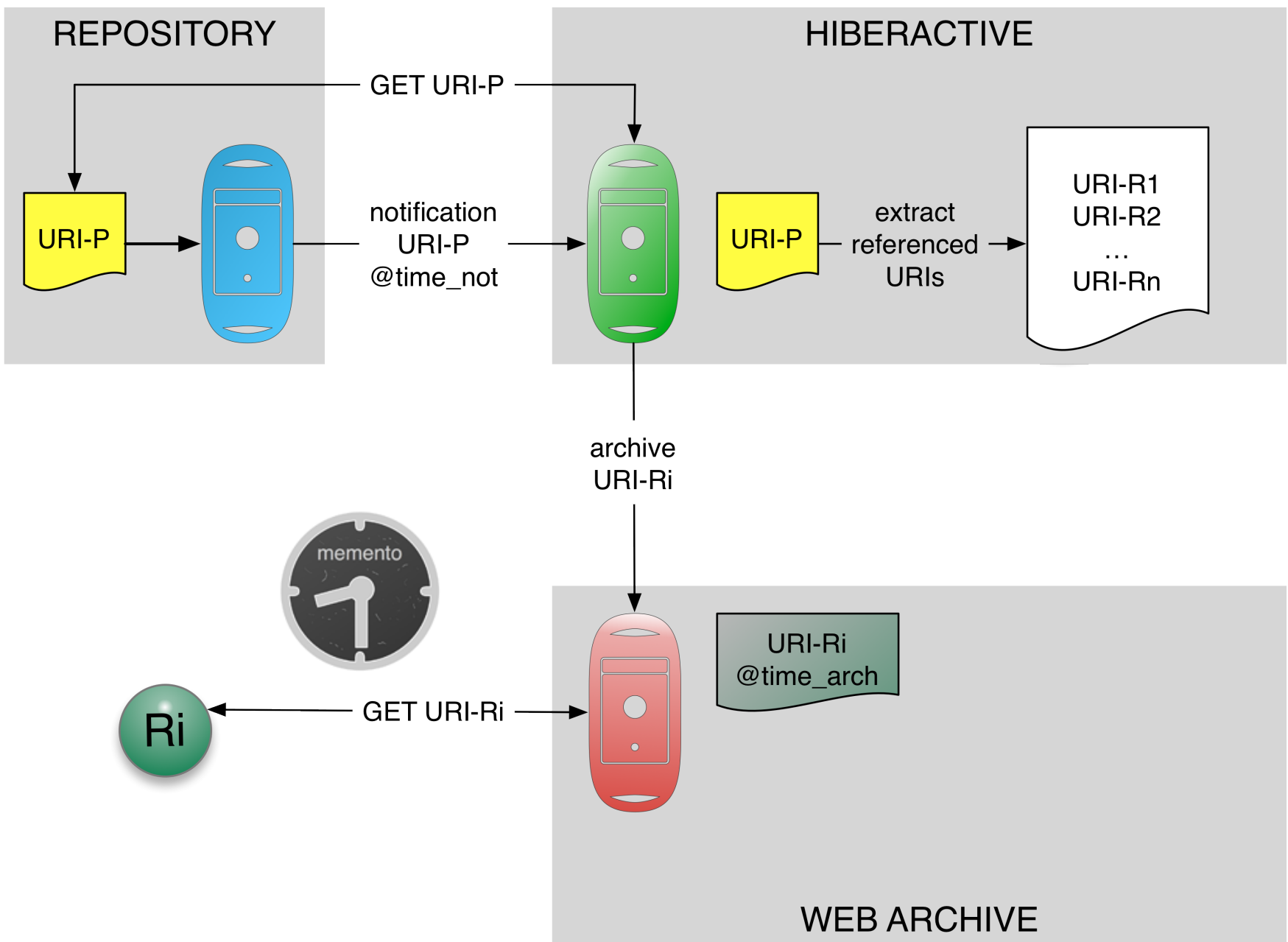
## Solving Reference Rot



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



**Los Alamos**  
NATIONAL LABORATORY  
THE UNIVERSITY of EDINBURGH



# 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