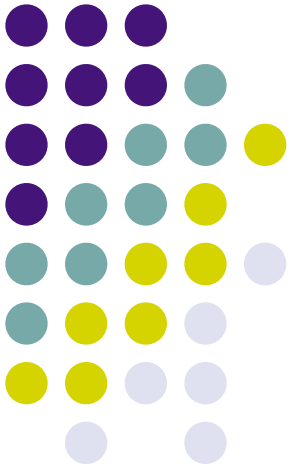
The Availability and Persistence of Web References in D-Lib Magazine

Frank McCown, Sheffan Chan,
Michael L. Nelson and Johan Bollen



IWAW05
September 22, 2005

Old Dominion University
Computer Science Department
Norfolk, Virginia, USA

D-Lib® Magazine

Search

Go

About D-Lib Magazine

Current Issue
- Table of Contents
- Featured Collection
- In Brief
- Clips & Pointers

Indexes
- Back Issues
- Author Index
- Title Index

Subscriptions

Search Guidelines

Mirror Sites

Author Guidelines

Contact D-Lib

DOI
10.1045/dlib.magazine

ISSN
1082-9873

RSS D-Lib via RSS

In the Current Issue

Full-length Features

July/August 2005
Vol. 11 No. 7/8

Also This Month

Digital Collections

FEATURED COLLECTION

Kinematic Models for Design Digital Library

An innovative digital library for learning and teaching about kinematics, the geometry of pure motion, and the history and theory of machines.

[Model from the Reuleaux Collection, Rolling Cones. Courtesy of Cornell University Library, KMODDL. Used with permission. Image cropped.]

Digital Library Community Activities

In Brief
Short items of current awareness.

In the News
Recent press releases and announcements.

Clips & Pointers
Documents, deadlines, calls for participation.

Archives

Back Issues and Indexes

References

Bekaert, Jeroen, Patrick Hochstenbach, and Herbert Van de Sompel. 2003. "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine*, Volume 9, Number 11, November 2003. <doi:10.1045/november2003-bekaert>.

Bekaert, Jeroen, Patrick Hochstenbach, Lyudmila Balakireva and Herbert Van de Sompel. 2004. "Using MPEG-21 and NISO OpenURL for the Dynamic Dissemination of Complex Digital Objects in the Los Alamos National Laboratory Digital Library,". *D-Lib Magazine*, Volume 10, Number 2, February 2004. <doi:10.1045/february2004-bekaert>.

Clausen, Lars. 2004. "Concerning Etags and Datestamps," Fourth International Web Archiving Workshop, ECDL 2004, Bath UK. <http://www.netarchive.dk/website/publications/Etags-2004.pdf>.

Freed, N. and N. Borenstein. 1996. "RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," November 1996. <http://www.ietf.org/rfc/rfc2045.txt?number=2045>.

Jerez, Henry, Xiaoming Liu, Patrick Hochstenbach, and Herbert Van de Sompel. 2004. "The multi-faceted use of the OAI-PMH in the LANL Repository," *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, June 7-11 2004, Tuscon, AZ, USA*. pp 11-20. <doi:10.1145/996350.996355>.

Kahn, Robert and Robert Wilensky. 1995. "A Framework for Distributed Digital Object Services. Corporation for National Research Initiatives," <http://www.cnri.reston.va.us/k-w.html>.

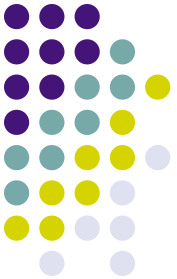
Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. "The Open Archives

Definition of Inaccessible URL

- Accessible URL: When performing an http GET on the URL, it should
  - return an http 200 (OK) response with non-zero length content

  OR
  - eventually return an http 200 response with non-zero length content after following one or more redirects (http 3xx)
- Inaccessible URL: Not an accessible URL (everything else)

Redirection Example



Request: http GET http://www.harding.edu/fmccown
Response: http 302 http://www.harding.edu/USER/fmccown/WWW

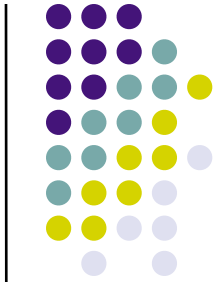
Request: http GET http://www.harding.edu/USER/fmccown/WWW
Response: http 301 http://www.harding.edu/USER/fmccown/WWW/

Request: http GET http://www.harding.edu/USER/fmccown/WWW/
Response: http 200 Content-Length: 765

Frequently encountered when using DOI resolvers, handles, and PURLs:

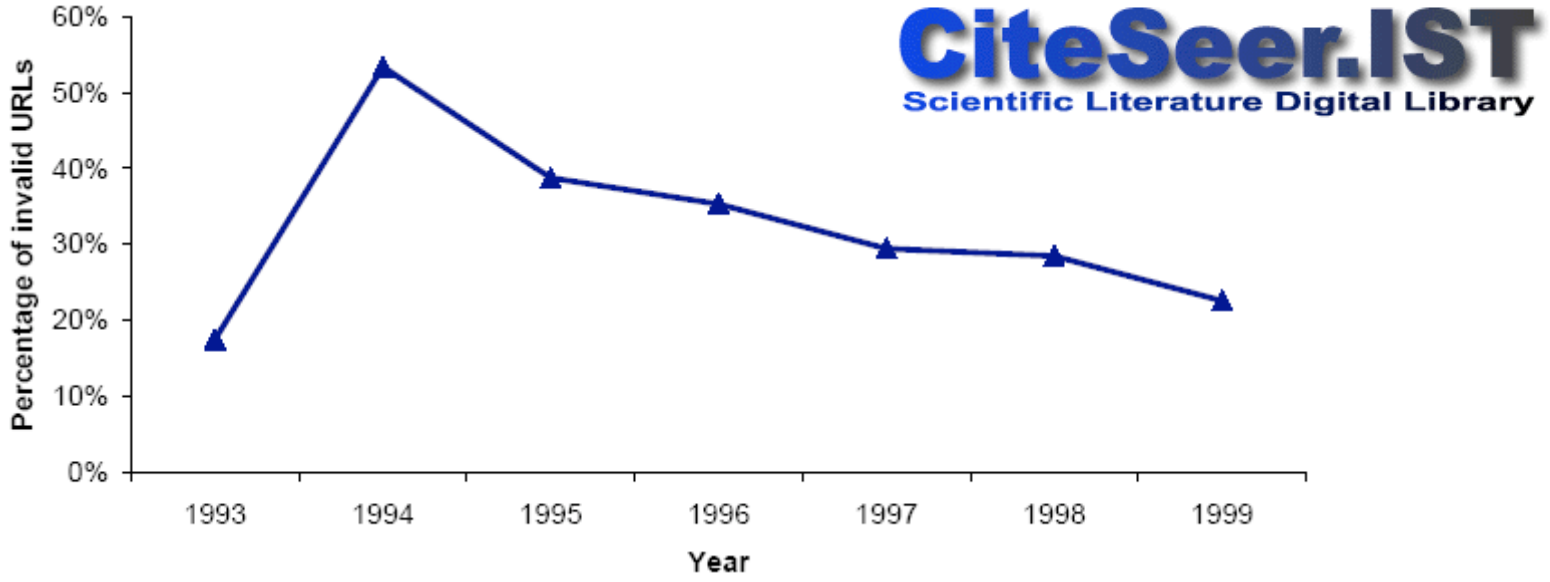
Request: http GET http://dx.doi.org/10.1045/april2001-liu
Response: http 302 http://www.dlib.org/dlib/april01/liu/04liu.html

Related Work



- Many papers discuss link-rot of academic citations
- 2 studies dealing with computer science and related articles
  - Steve Lawrence et al., “Persistence of Web References in Scientific Research”, IEEE Computer, 34(2), 2001
  - Diomidis Spinellis, “The Decay and Failures of Web References”, Communications of the ACM, 46(1), 2003

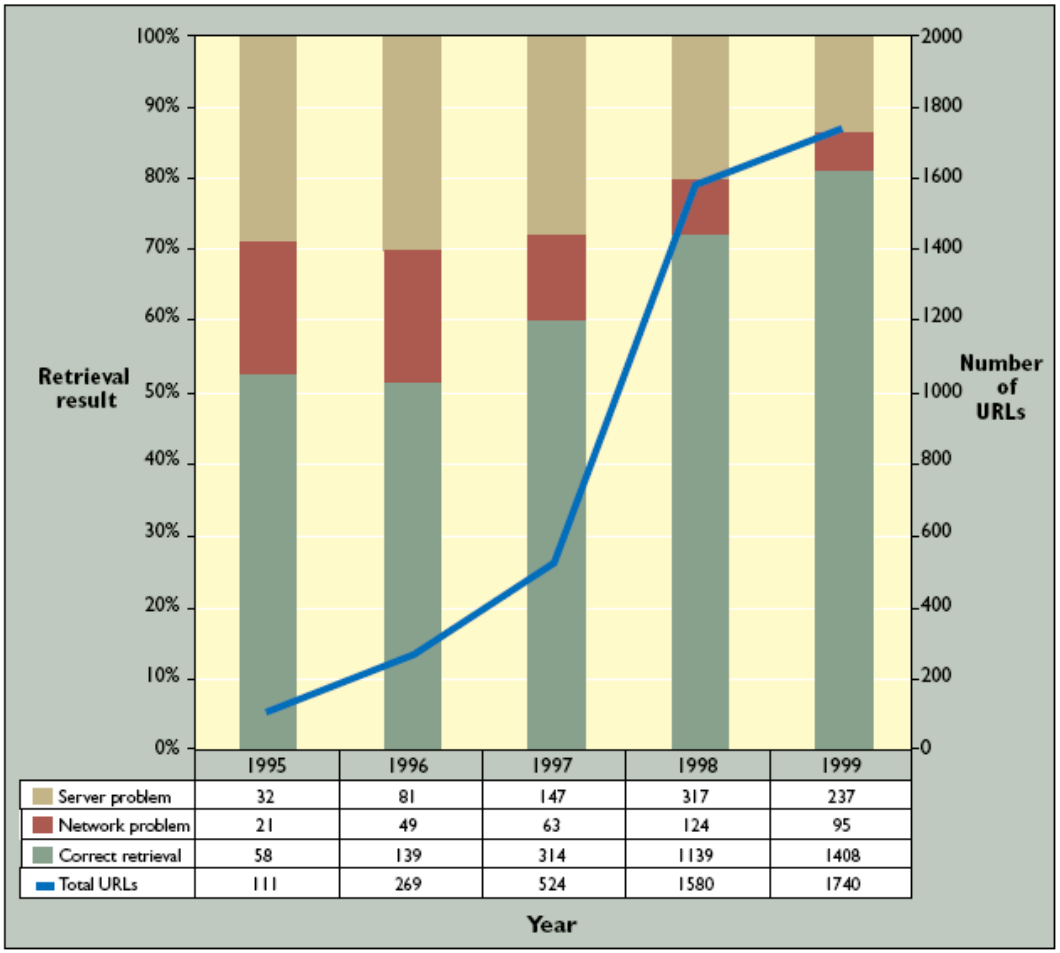
Lawrence Study



67,577 URLs accessed in May 2000
Half-life of URL = 6 years from publication date (our calculation)

Figure from http://www.searchlores.org/library/persistence-computer01.pdf

Spinellis Study

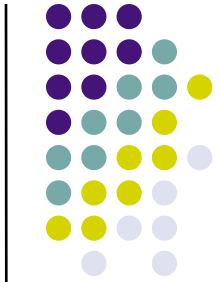


	1995	1996	1997	1998	1999
Server problem	32	81	147	317	237
Network problem	21	49	63	124	95
Correct retrieval	58	139	314	1139	1408
Total URLs	111	269	524	1580	1740

- 1,391 URLs from *Communications of the ACM*
- 2,833 URLs from *IEEE Computer*
- Accessed in June 2000
- Half-life of URL = 4 years from publication date

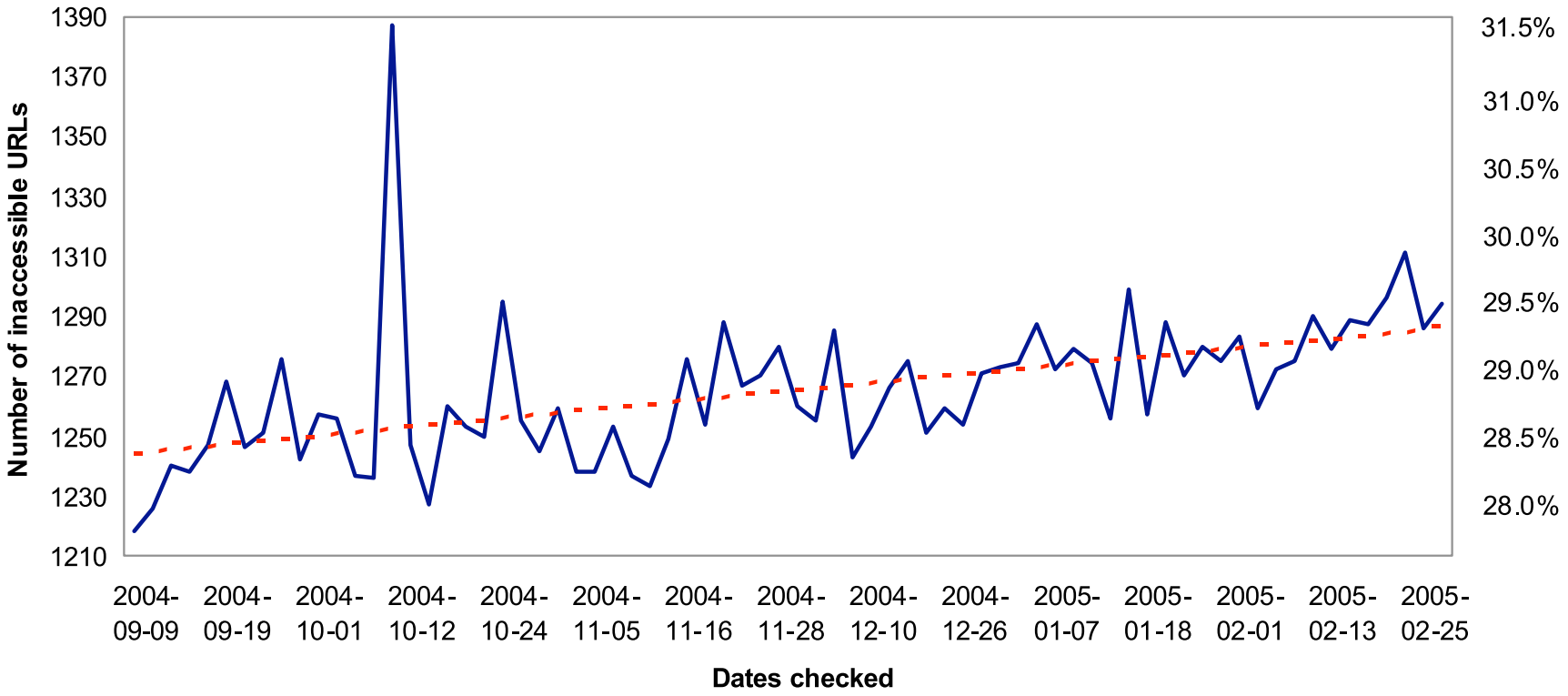
Figure from http://portal.acm.org/citation.cfm?doid=602421.602422

Methodology

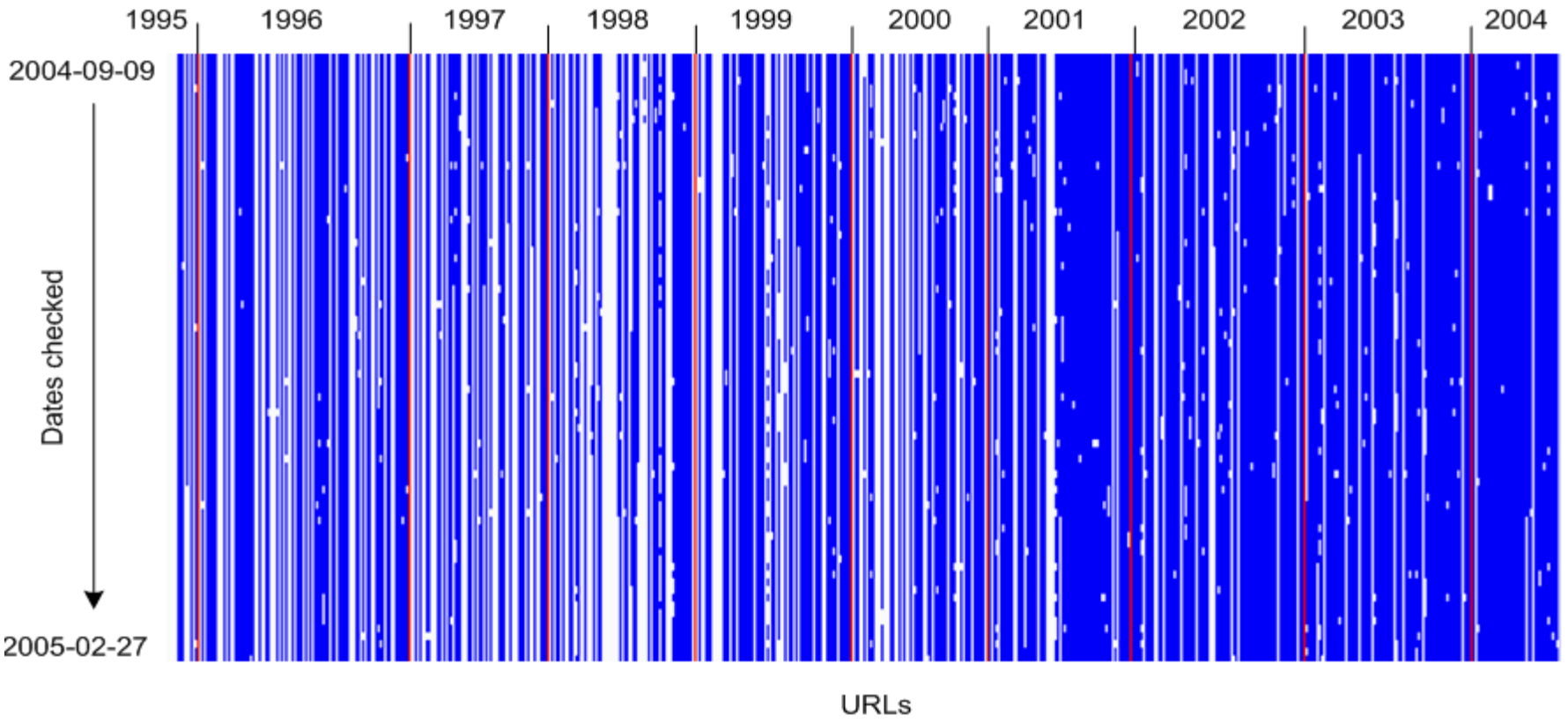


1. Downloaded all articles from July 1999 to August 2004 (453 articles) and extracted all hyperlinks (7094 total).
2. Removed all URLs that referenced www.dlib.org (http://dx.doi.org/10.1045/* and http://www.dlib.org/*) and all redundant URLs, producing a total of 4387 URLs
3. Downloaded 4387 URLs 72 times (three times a week for 25 weeks), beginning on September 9, 2004 and ending on February 27, 2005

Availability at Checkpoints

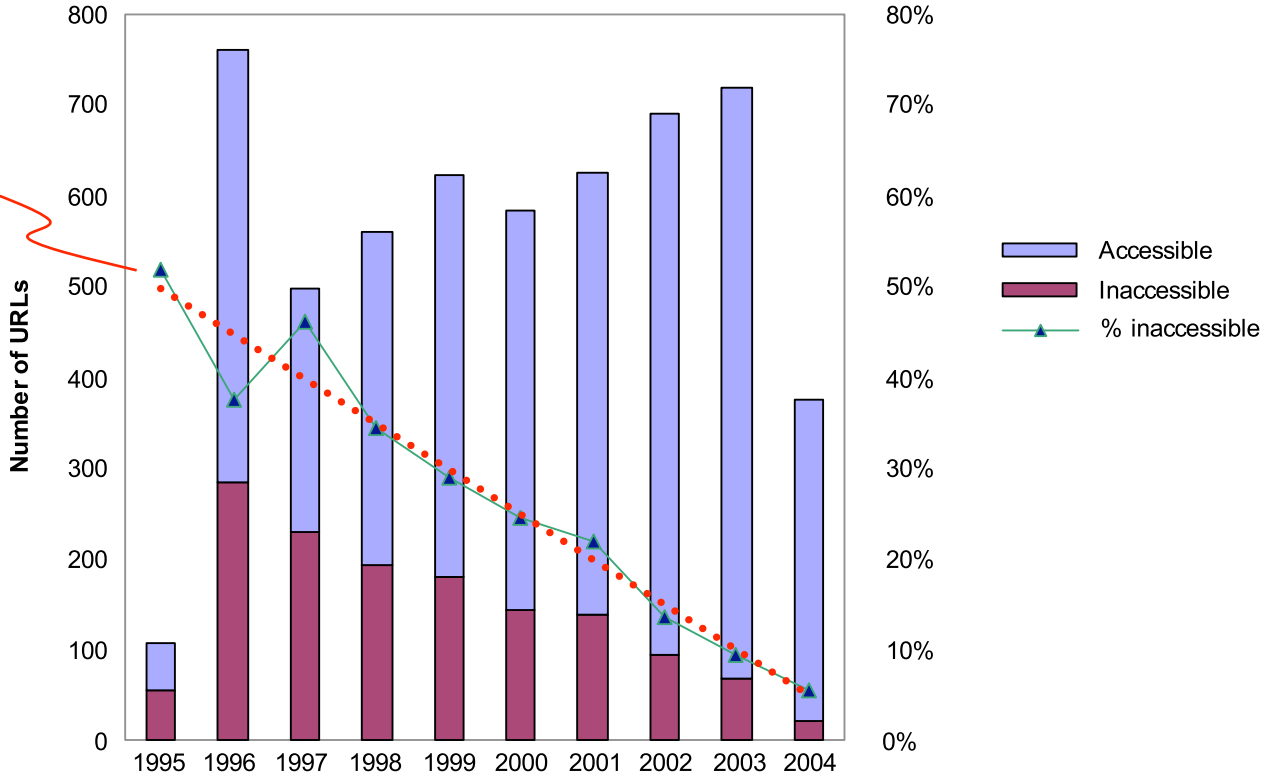


Availability at Checkpoints



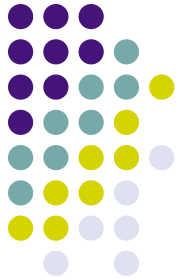
Distribution by Year

10 year half-life from publication date



	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Total URLs	106	761	498	560	624	584	626	690	719	376
Total Articles	19	55	59	51	50	48	45	49	52	25
URLs per article	5.6	13.8	8.4	11	12.5	12.2	13.9	14.1	13.8	15

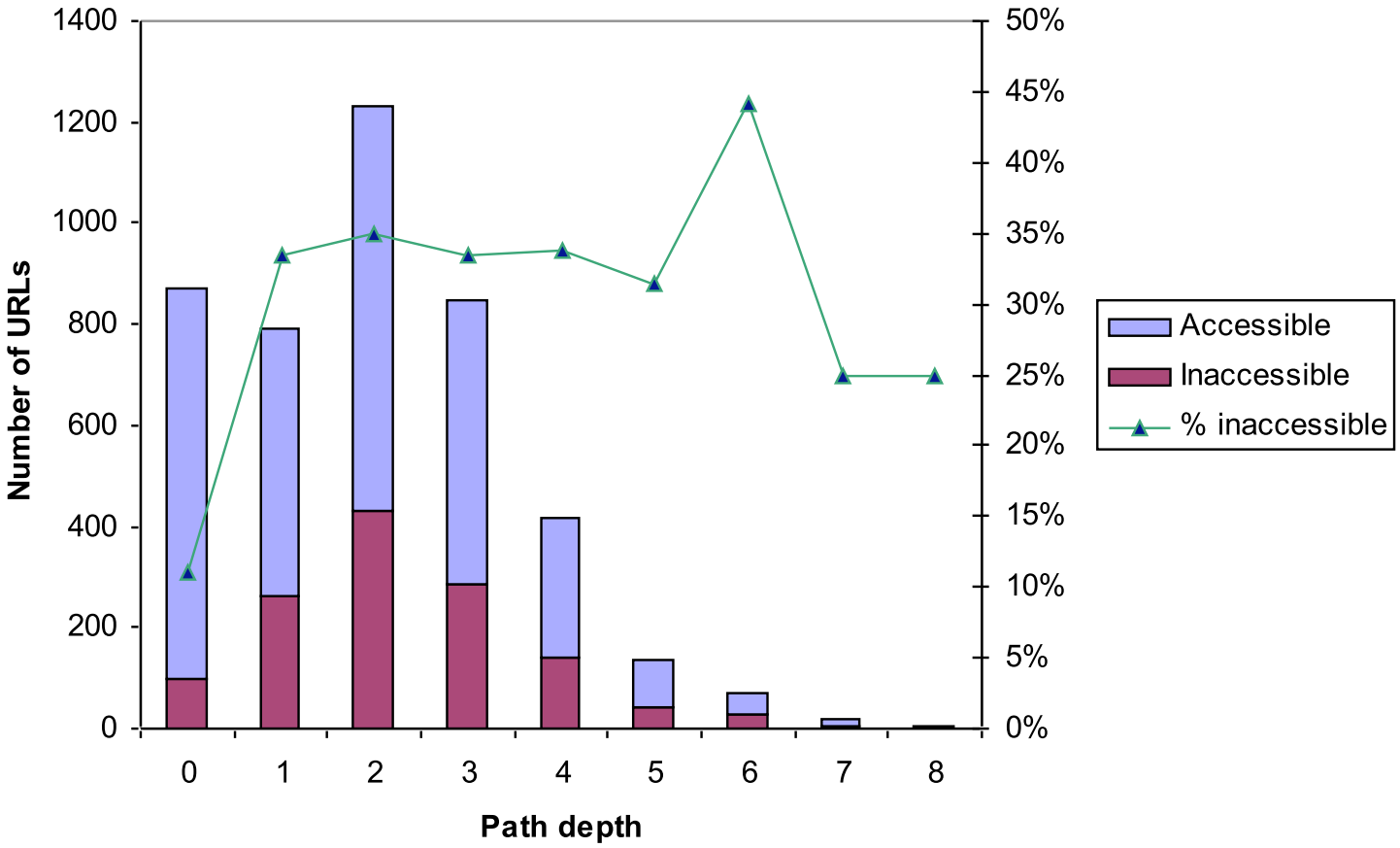
Error Codes



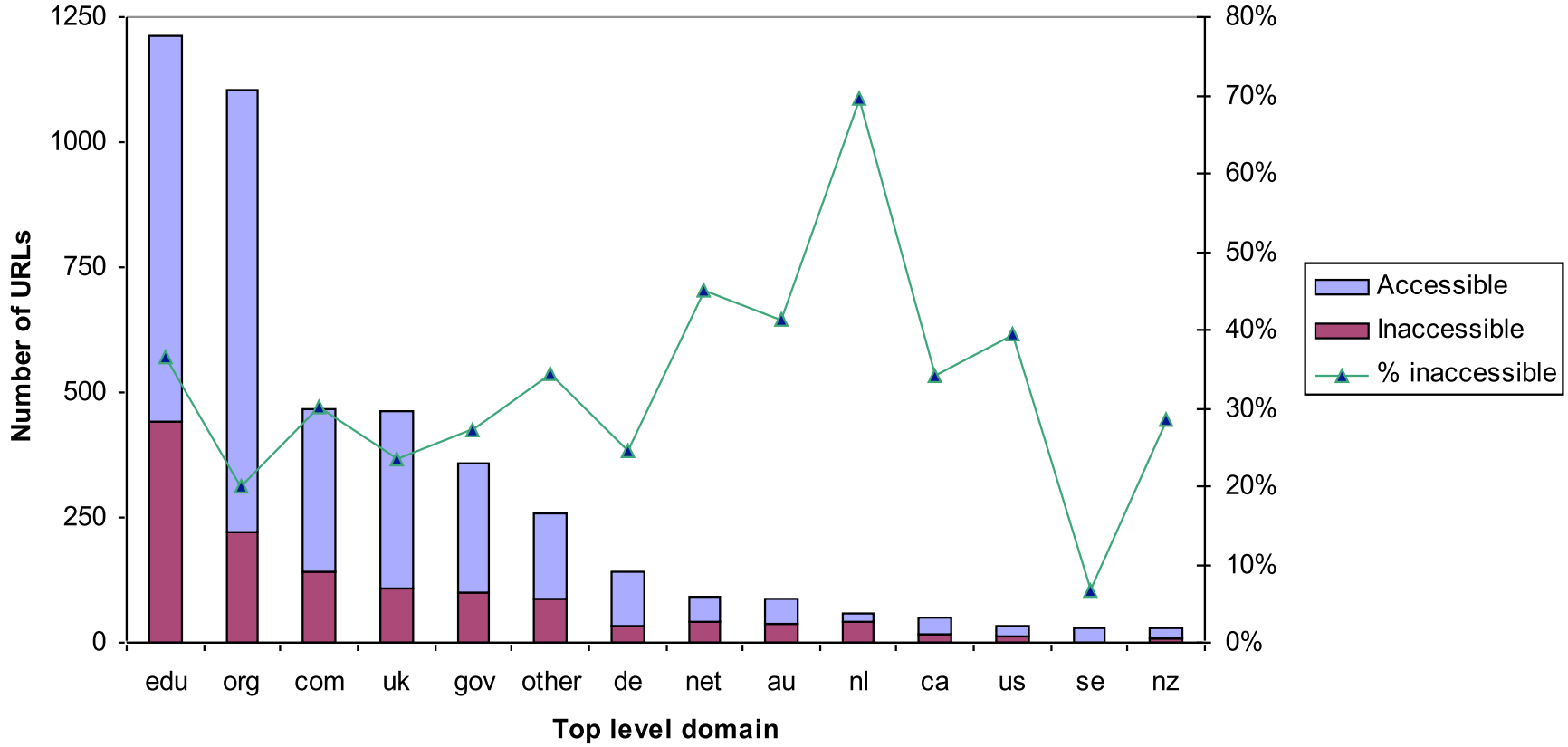
HTTP Code	Meaning	First check	Last check
404	Not found	62.40 %	60.20 %
500	Internal sever error	32.51%	35.09 %
403	Forbidden	3.94 %	3.86 %
401	Unauthorized	0.74 %	0.62 %
200	OK but 0 length content	0.25 %	0.23 %
410	Gone	0.08 %	0.00 %
502	Bad gateway	0.08 %	0.00 %

“Soft 404s” were not tested.

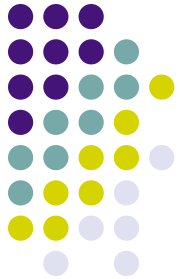
Path Depth


1400
1200
1000
800
600
400
200
0
Number of URLs
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0
1
2
3
4
5
6
7
8
Path depth
Accessible
Inaccessible
% inaccessible

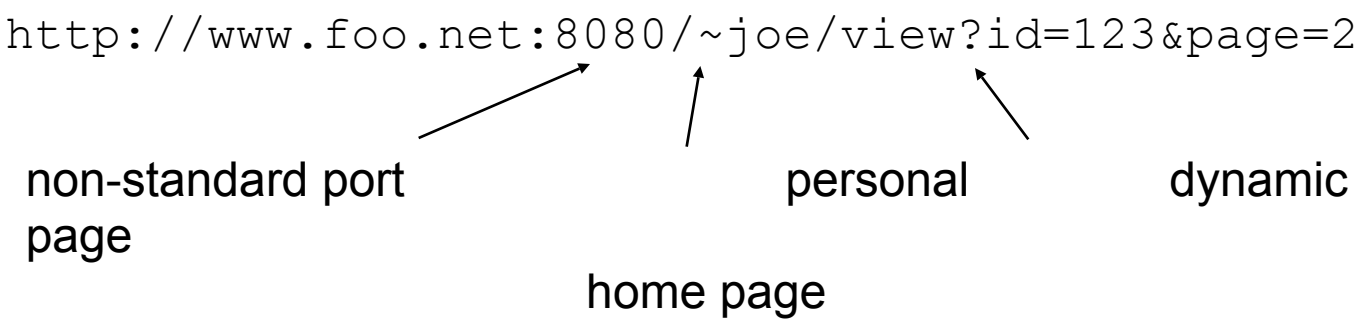
Top-Level Domain



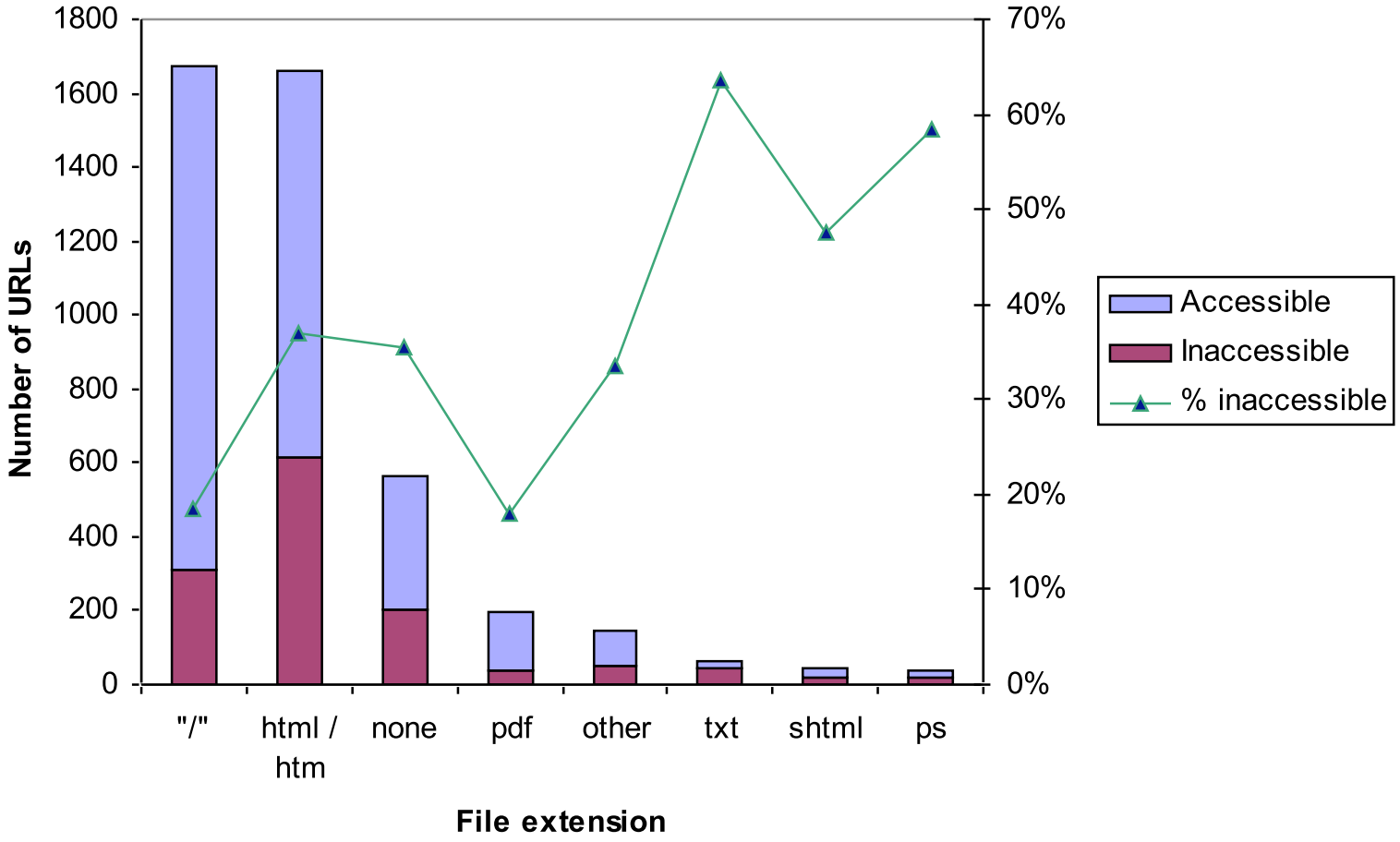
Path Characteristics



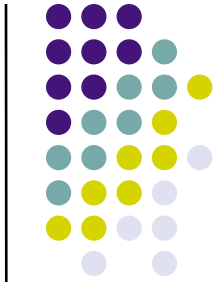
	Personal home page	Non-standard port	Dynamic page
Inaccessible URLs	136	53	76
Accessible URLs	126	11	109
Total URLs	262	64	185
% Inaccessible	51.9 %	82.8 %	41.1 %



File Extension

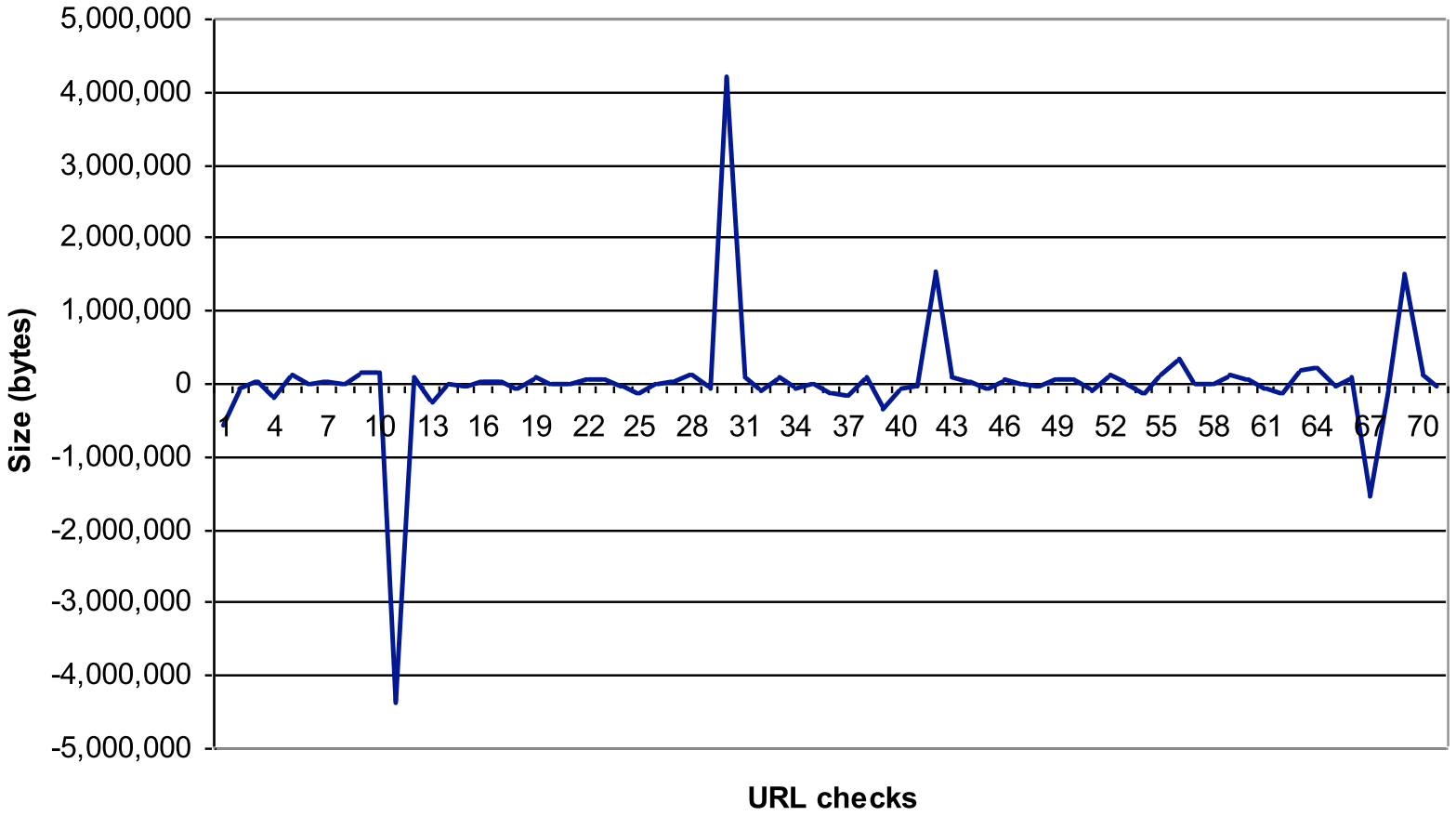


Persistent URLs



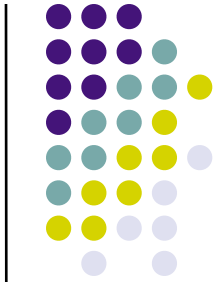
- Few uses of mechanisms designed to make URLs persist
- 59 PURLs (unique) - 59% were inaccessible
- 2 handles (unique) – none inaccessible
- 15 DOIs (unique, not pointing back to dlib.org) – none inaccessible

Content Changes



683 “In flux” URLs → More than 1KB change in size

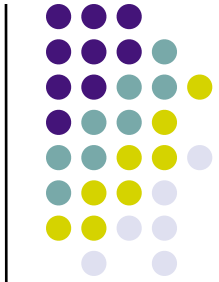
Bad URL Characteristics



The URL characteristics below were associated with increased levels of linkrot:

- a non-standard port
- a personal homepage
- dynamic query strings
- uncommon or deprecated file extensions (e.g., .txt, .shtml, .ps)
- .net, .edu or country-specific top-level domain names

Thank You



Questions?

Slides and data files:

http://www.cs.odu.edu/~fmccown/research/dlib_urls/