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

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

Old Dominion University
Computer Science Department
Norfolk, Virginia, USA

September 22, 2005
In the Current Issue

Full-length Features

July/August 2005
Vol. 11 No. 7/8
Table of Contents

EDITORIAL
Ten Years of D-Lib Magazine and Counting
by Robert E. Kahn, CNRI

LETTERS
To the Editor

ARTICLES
A Tenth Anniversary for D-Lib Magazine
by Bonita Wilson and Allison L. Powell,
Corporation for National Research Initiatives
doi:10.1045/july2005-wilson

Really 10 Years Old?
by Amy Friedlander, Shinkuro, Inc.
doi:10.1045/july2005-friedlander

Whence Leadership?
by Ronald L. Larsen, University of Pittsburgh
doi:10.1045/july2005-larsen

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

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
References


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
  - 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
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

- 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
## Error Codes

<table>
<thead>
<tr>
<th>HTTP Code</th>
<th>Meaning</th>
<th>First check</th>
<th>Last check</th>
</tr>
</thead>
<tbody>
<tr>
<td>404</td>
<td>Not found</td>
<td>62.40 %</td>
<td>60.20 %</td>
</tr>
<tr>
<td>500</td>
<td>Internal server error</td>
<td>32.51 %</td>
<td>35.09 %</td>
</tr>
<tr>
<td>403</td>
<td>Forbidden</td>
<td>3.94 %</td>
<td>3.86 %</td>
</tr>
<tr>
<td>401</td>
<td>Unauthorized</td>
<td>0.74 %</td>
<td>0.62 %</td>
</tr>
<tr>
<td>200</td>
<td>OK but 0 length content</td>
<td>0.25 %</td>
<td>0.23 %</td>
</tr>
<tr>
<td>410</td>
<td>Gone</td>
<td>0.08 %</td>
<td>0.00 %</td>
</tr>
<tr>
<td>502</td>
<td>Bad gateway</td>
<td>0.08 %</td>
<td>0.00 %</td>
</tr>
</tbody>
</table>

“Soft 404s” were not tested.
Path Depth

![Path Depth Graph]

- **Accessible**
- **Inaccessible**
- **% Inaccessible**

**Path Depth**

Number of URLs vs Path depth

- 0 Path Depth: 800 URLs
- 1 Path Depth: 800 URLs
- 2 Path Depth: 1200 URLs
- 3 Path Depth: 800 URLs
- 4 Path Depth: 300 URLs
- 5 to 8 Path Depth: Decreasing number of URLs
Top-Level Domain
## Path Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Personal home page</th>
<th>Non-standard port</th>
<th>Dynamic page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inaccessible URLs</td>
<td>136</td>
<td>53</td>
<td>76</td>
</tr>
<tr>
<td>Accessible URLs</td>
<td>126</td>
<td>11</td>
<td>109</td>
</tr>
<tr>
<td>Total URLs</td>
<td>262</td>
<td>64</td>
<td>185</td>
</tr>
<tr>
<td>% Inaccessible</td>
<td>51.9 %</td>
<td>82.8 %</td>
<td>41.1 %</td>
</tr>
</tbody>
</table>

The given URL is:

```
```
File Extension

The diagram illustrates the number of URLs for various file extensions and the percentage of inaccessible URLs. The x-axis represents file extensions, and the y-axis shows the number of URLs. The graph includes a line indicating the percentage of inaccessible URLs.
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/