


2002

Object Persistence and Availability in Digital Libraries

Michael L. Nelson
Old Dominion University

B. Danette Allen

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_fac_pubs

 Part of the [Computer Sciences Commons](#), and the [Digital Communications and Networking Commons](#)

Repository Citation

Nelson, Michael L. and Allen, B. Danette, "Object Persistence and Availability in Digital Libraries" (2002). *Computer Science Faculty Publications*. 2.

https://digitalcommons.odu.edu/computerscience_fac_pubs/2

Original Publication Citation

Nelson, M.L., & Allen, B.D. (2002). Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1), 1-24. doi: 10.1045/january2002-nelson

ARTICLES

D-Lib Magazine
January 2002

Volume 8 Number 1

ISSN 1082-9873

Object Persistence and Availability in Digital Libraries[Michael L. Nelson](#), [B. Danette Allen](#)

NASA Langley Research Center

Hampton, VA 23681

{m.l.nelson, b.d.allen}@nasa.gov

Abstract

We have studied object persistence and availability of 1,000 digital library (DL) objects. Twenty World Wide Web accessible DLs were chosen and from each DL, 50 objects were chosen at random. A script checked the availability of each object three times a week for just over 1 year for a total of 161 data samples. During this time span, we found 31 objects (3% of the total) that appear to no longer be available: 24 from PubMed Central, 5 from IDEAS, 1 from CogPrints, and 1 from ETD.

1.0 Introduction

We have measured the persistence and availability of digital objects in a variety of publicly available World Wide Web (WWW) digital libraries (DLs). We selected twenty DLs, and from those DLs selected 50 information objects, for a total of 1,000 information objects. Three times a week (Sunday, Tuesday and Friday early morning, Eastern Time zone) from November 21, 2000, through December 9, 2001, scripts were run that download the information object and record the number of bytes successfully transmitted. All tests were run from the machine ruby.ils.unc.edu (152.2.81.1; Solaris 2.7). When the terms "unavailable" and "inaccessible" are used below, it is important to remember that these terms are with respect to ruby.ils.unc.edu.

We are particularly interested in the long-term availability of individual objects in the DL. Similar previous studies have looked at the availability of general http servers ([Viles & French, 1995](#)), or at the availability of general web pages ([Douglass et al., 1997](#); [Koehler, 1999](#); [Koehler, 2000](#)). One study looked at the availability of DL services that were layered on top of http ([Powell & French, 2000](#)). We find similar network and http server availability issues that these studies report. However, for evaluating the availability of individual objects, these studies are insufficient. Certainly, if the http server, DL service, or "entry" web page is missing, then the desired object will be unreachable. However, simply verifying the operation of an http server, DL service or web page does not imply that the actual object itself is still available or unchanged. On the assumption that being placed in a DL is indicative of someone's desire to increase the persistence and availability of an object, we expect DL objects to survive longer, change less, and be more available than general WWW content.

Of the objects that appeared to be unavailable from the test data, we manually sought out the objects in the tested DL, and in the case of distributed DLs, we went to the original web sites. After doing

so, we consider 31 of the 1,000 objects to be currently unavailable. It is possible that these objects are available in some other form, or by another name, and we were simply unable to find them. However, these 31 objects represent results returned in searching and browsing of the DLs in November 2000 that we cannot currently find. This percentage is similar to what Lawrence et al. (2001) report in their study of the persistence of URLs that appear in technical papers. For recently authored papers, they found over 20% of the URLs listed were invalid. However, manual searches were able to reduce the number of "lost" URLs to 3%. Since not all the URLs extracted for their study were necessarily objects in DLs, we are unsure if 3% lossage is generalizable.

2.0 Test Collection

The twenty DLs we studied are listed in Table 1. We selected these DLs for their mixture of coverage, popularity, geographic locations, and data storage architectures. Some of the DLs have their own "brand" or "identity", some are part of an institution's library services, and others are simply report listings on a web page. However, we considered them to be DLs regardless of their own awareness or promotion of themselves as DLs. "Contributor" describes from where the DL's content comes, and "Access" indicates if the URLs for the objects represent direct access to the delivery format (i.e., PDF, PostScript) or if access goes through a "wrapper" service that delivers the content through a server-side process in addition to the http server. The appendix contains a tar file of all the objects, scripts used to test the objects, the test data, and the Matlab scripts used to generate the visualization of the data.

DL Root	Content	Data Storage	Contributor	Access
w.harvard.edu	Astrophysics	Centralized & distributed	Publishers, Societies, Universities	Wrapper
ccdb.kek.jp	Astrophysics	Centralized	Universities, Laboratories	Wrapper
citeseer.nj.nec.com	Computer science	Centralized & distributed	Scraped From Web Pages	Wrapper
cogprints.soton.ac.uk	Cognitive science	Centralized	Individual Authors	Direct
data.mpi-sb.mpg.de	Mathematics, computer science	Centralized	MPI	Wrapper
ideas.uqam.ca	Economics	Distributed	Individual authors	Direct
mtrs.msfc.nasa.gov	Aerospace	Centralized	MSFC	Direct
pubmedcentral.nih.gov	Medical science	Centralized	Publishers	Wrapper
stinet.dtic.mil	General science	Centralized	Universities, Laboratories	Wrapper
www.arxiv.org	Physics, mathematics, computer science	Centralized	Individual authors	Wrapper
www.inria.fr	Computer science	Centralized	INRIA	Direct
www.jlab.org	Physics	Centralized	JLab	Direct
www.ncstrl.org	Computer science	Distributed	Universities	Wrapper
www.nzdl.org	Computer science	Centralized & distributed	Universities	Wrapper

www.onera.fr	Aerospace	Centralized	ONERA	Direct
www.rand.org	Economics, mathematics, social sciences	Centralized	RAND	Direct
www.research.digital.com	Computer science	Centralized	COMPAQ/DEC	Direct
www.santafe.edu	Mathematics, economics	Centralized	SFI	Direct
www.slac.stanford.edu	Physics	Centralized	Universities, Laboratories	Direct
www.theses.org	General science	Centralized	Universities	Direct

Table 1: DLs Tested

All twenty DLs serve reports, e-prints, or re-prints of scientific and technical information. Some were author contributed (e.g., arXiv, CogPrints); some held organizationally produced report series (e.g., Santa Fe, ONERA) and some collect their holdings by "spidering" known DLs (e.g., CiteSeer). Objects were collected by performing searches on broad terms to produce large numbers of hits (e.g., "systems", "complex"). For the few DLs that did not offer searching, we selected objects from their report listings to cover the range of their offerings. Where possible, we linked to the PDF or PostScript versions of the reports, although some were available only as ASCII, HTML or GIF. For arXiv, we linked directly to the TeX source files to avoid generating a dynamic conversion to another format. For CiteSeer, we linked to the copy in their local cache, and not the remote copy. All the objects were stored as URLs; none of the DLs directly used any URN implementation (e.g., Purls, Handles, or DOIs) with globally registered names.

3.0 Results

There were 161 test days and 1,000 objects for a total of 161,000 download opportunities. The individual results are listed in Table 2. In summary, DLs were completely unreachable 2% of the time; individual objects were unavailable 11% of the time; and the objects changed in size more than one kilobyte 5% of the time and had at least some change from their baseline 22% of the time. Four test days were thrown out (2001-04-01, 2001-06-24, 2001-07-29, 2001-09-18) because the testing program on the host machine did not run to completion on those days.

DL Root	Days Entire DL Not Available	Failed Object Accesses	> 1KB Size Change	Any Size Change
		(N-161000)		
adswww.harvard.edu	0	194	1	3919
ccdb.kek.jp	0	372	2383	7343
citeseer.nj.nec.com	0	5550	281	5880
cogprints.soton.ac.uk	1	91	0	91
data.mpi-sb.mpg.de	5	251	0	251
ideas.uqam.ca	0	312	55	609
mtrs.msfc.nasa.gov	11	553	38	591

pubmedcentral.nih.gov	1	3562	1	3873
stinet.dtic.mil	7	494	12	806
www.arxiv.org	1	89	512	700
www.inria.fr	2	481	0	481
www.jlab.org	0	170	1228	1398
www.ncstrl.org	0	2884	706	3675
www.nzdl.org	12	615	0	615
www.onera.fr	10	514	50	564
www.rand.org	1	577	2061	3643
www.research.digital.com	2	148	0	148
www.santafe.edu	1	76	0	293
www.slac.stanford.edu	0	0	0	0
www.theses.org	1	49	157	206
Totals:	55	16982	7485	35086
Percentage:	2%	11%	5%	22%

Table 2. Results (2000-11-21 through 2001-12-09)

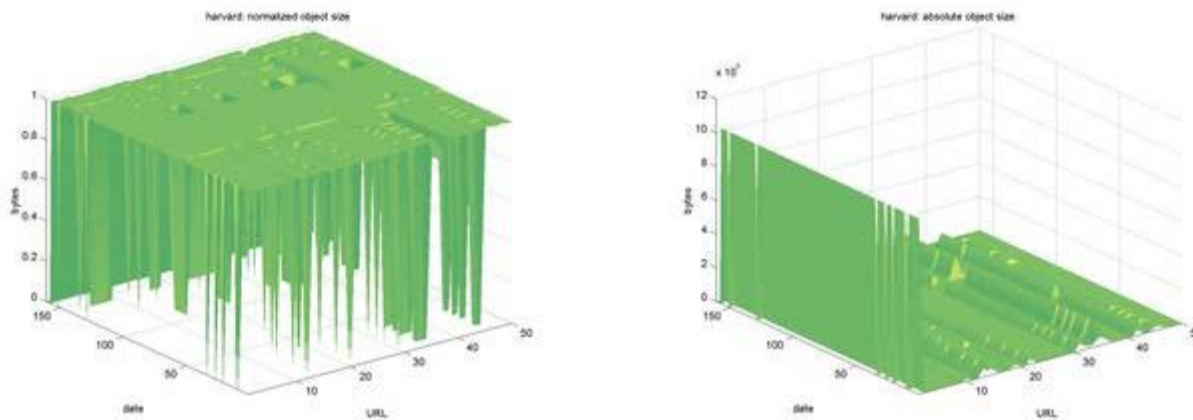
When viewing the figures below, it is important to keep in mind that the details are less important than the overall trends exhibited. In particular, the figures showing the normalized byte count from the DLs can be thought of a floor or a walking surface. Goodness is represented by a smooth, even surface and a surface that would appear treacherous to walk is indicative of a DL with suspect availability.

3.1 Astrophysics Data System (adswww.harvard.edu)

The Astrophysics Data System (ADS) <<http://adswww.harvard.edu>> is a NASA-funded project that collects abstracts and articles in astrophysics, physics, planetary science, astronomy, and related disciplines. The contents are pulled from many sources, including: NASA holdings, commercial publishers, professional societies, and university theses and dissertations. Some of the content is held locally in the ADS system, and some content is held remotely (the contents of commercial serials typically are held at the publishers' sites). Some of the reports are served as PDFs only (generally, these are served from the publisher's site), and other reports are served with their default as paginated GIF files, with PDF and other options and services available. All objects are accessed through a resolver service and specified by their bibcode (a unique identifier for publications used within the astrophysics community). An example object URL is:

http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000LPI...31.1903M&link_type=ARTICLE&db_key=AST

Figure 1 shows the results for ADS. There is some sporadic unavailability of some of the objects, but these appear to be the result of "bad network days". Since not all objects are stored centrally at ADS, there appear to be "pockets" of inaccessible objects for certain days, but this does not apply to the entire collection. The size varies for some of the objects, but this is most likely due to many of the objects having wrapper pages. Any change in the wrapper page will yield a different byte size. In summary, all objects remain available.



Astrophysics Data System (normalized byte count) Astrophysics Data System (absolute byte count)

[MPEG Rotation](#)

[MPEG Rotation](#)

Figure 1. ADS Results

3.2 KISS (KEK Information Service System) for Preprints (ccdb.kek.jp)

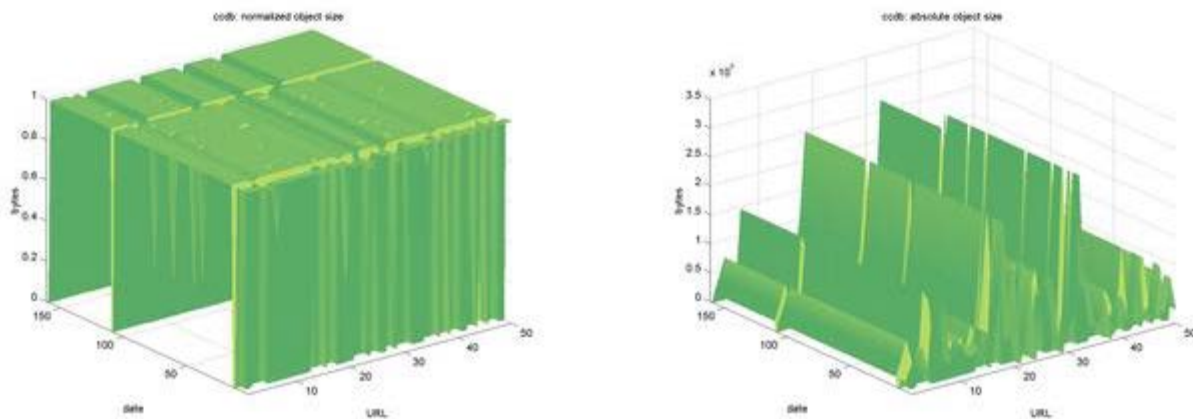
The KISS server holds preprints from the High Energy Accelerator Research Organization (KEK) library in Tsukuba, Japan. The preprint search interface is available from a slightly different URL <http://www-lib.kek.jp/KISS/kiss_prepri.html> than that of the documents themselves. The preprint library contains scans of paper preprints from other high-energy physics laboratories. The objects are available in TIFF, GIF and PostScript from URLs similar to:

http://ccdb.kek.jp/cgi-bin/img_index?196700118

Where the last part of the URL is the KEK accession number. For this study, we explicitly chose the Level 2 Unix compressed PostScript option, with URLs similar to:

<http://ccdb.kek.jp/cgi-bin/img/tiff2allps?2+compress+196700118>

Figure 2 shows the results for KISS. Other than 2 periods of network unavailability (2001-12-15, 2001-12-17, 2001-12-19; and 2001-08-05, 2001-08-07), the objects remained highly available. KISS did produce many small changes in the number of bytes returned, the reason for which remains unknown. All objects in KISS remain available.



KEK Information Service System for Preprints
(normalized byte count)

[MPEG Rotation](#)

KEK Information Service System for Preprints
(absolute byte count)

[MPEG Rotation](#)

Figure 2. KISS Results

3.3 ResearchIndex (citeseer.nj.nec.com)

ResearchIndex is a demonstration DL of NEC's Autonomous Citation Indexing (ACI) software. Although the ACI software can be used for a DL of any discipline, the ResearchIndex is currently focused on computer science. ResearchIndex has a robot that crawls known sites, looking for PDF and PostScript versions of reports and pre-prints. It then automatically extracts the citations from the file, and builds the indices according to reference patterns in the collection. Since it draws its objects from remote sites, ResearchIndex also caches the objects centrally. It provides users with the option to download from the original site or use the version cached centrally.

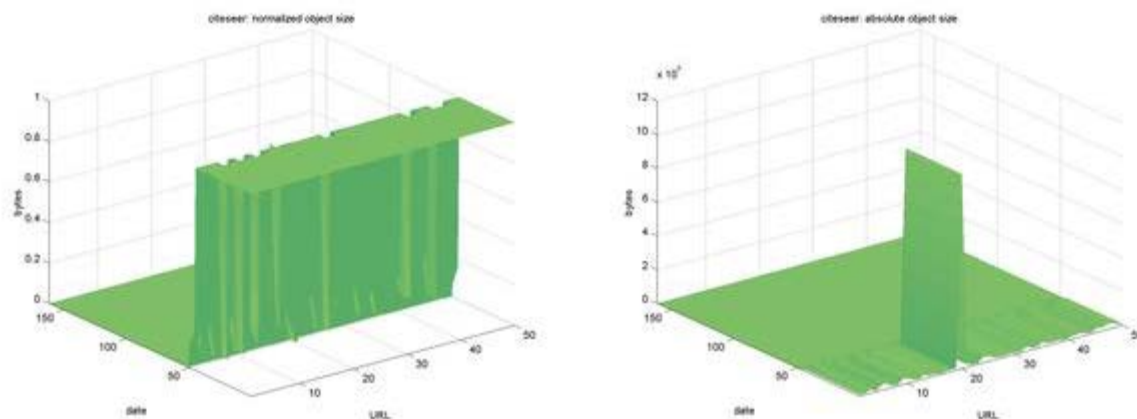
Figure 3 shows the results for ResearchIndex. At first glance, it appears that all of the objects are no longer accessible. However, it turns out that the API for ResearchIndex has changed. We choose all 50 objects to use the centrally cached version. The URLs originally looked like:

```
http://citeseer.nj.nec.com/cs?profile=0%2C350614%2C1%2C0.25%2CDownload
&rd=http%3A//citeseer.nj.nec.com/papers/cs/2883/http%253A%252
3@S@%2523%2523@S@%2523www.cs.tu-berlin.de%2523@S@%2523%257Etolk%2
523@S@%2523linda%2523@S@%2523Open%2523@S@%2523UIUCDCS-R-92-1766.ps.gz
```

However, starting on 2001-03-06 the byte sizes of all the objects fall to similar sizes (approximately 11,000 bytes). Then, by 2001-03-18 all the file sizes are zero. Presumably, the former dates had a warning page indicating the change. URLs for the objects are now of the form:

```
http://citeseer.nj.nec.com/rd/28106572%2C350614%2C1%2C0.25%2CDownload
/http%253A%252F%252Fciteseer.nj.nec.com/cache/papers/cs/2883/http%253Az
SzzSzwww.cs.tu-berlin.dezSz%257EtolkzSzlindazSzOpnzSzUIUCDCS-R-92-
1766.ps.gz/actorspaces-an-open-distributed.ps.gz
```

All 50 objects were manually found in ResearchIndex, even though their URLs changed over time.



ResearchIndex (normalized byte count)

[MPEG Rotation](#)

ResearchIndex (absolute byte count)

[MPEG Rotation](#)

Figure 3. ResearchIndex Results

3.4 CogPrints (cogprints.soton.ac.uk)

CogPrints is an author contributed e-print archive that covers cognitive science and related disciplines. CogPrints accepts many different formats from authors, including HTML, PostScript and PDF. Objects have URLs similar to:

<http://cogprints.soton.ac.uk/documents/disk0/00/00/05/37/cog00000537-00/Cdt.ps>

Figure 4 shows the results for CogPrints. The entire DL was unreachable on 2001-09-21. There are four objects shown on the graph that eventually became unavailable. The first object became obsolete during the time between when the initial objects were chosen and the time the baseline was run. The object:

http://cogprints.soton.ac.uk/documents/disk0/00/00/05/47/cog00000547-00/Neurocomp98_Chaos.ps.Z

no longer exists, but CogPrints redirects you to the object for which it believes you are looking. In this case, the compressed version of this PostScript file is no longer available, and deleting the trailing ".Z" is sufficient to regain access to this object.

Two objects:

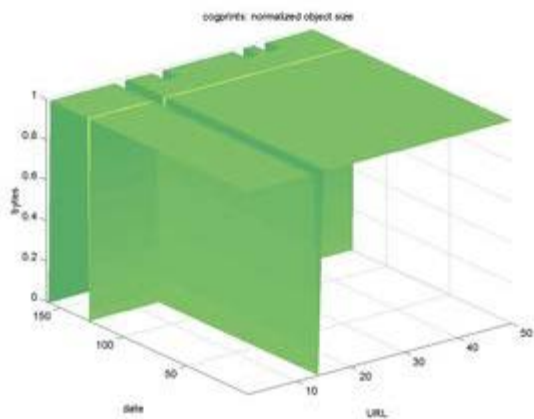
<http://cogprints.soton.ac.uk/documents/disk0/00/00/04/09/cog00000409-00/tokyo99.htm>

<http://cogprints.soton.ac.uk/documents/disk0/00/00/07/95/cog00000795-00/ACCS.htm>

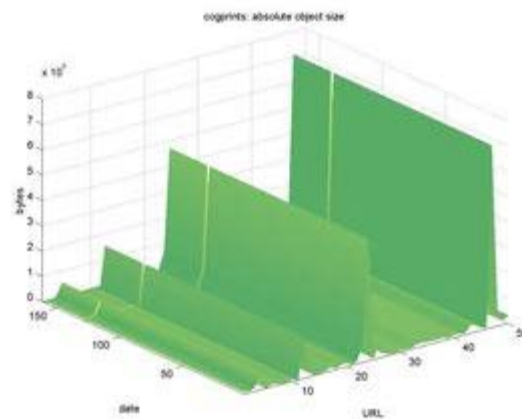
eventually became unavailable, but a manual check reveals that these e-prints were removed. Pointers to later versions are provided. However, there remains one object:

<http://cogprints.soton.ac.uk/documents/disk0/00/00/08/75/cog00000875-00/ACCS.htm>

that was removed for which no pointer to a later version was given. It is unknown if this object was either accidentally lost, or intentionally removed.



CogPrints (normalized byte count)
[MPEG Rotation](#)



CogPrints (absolute byte count)
[MPEG Rotation](#)

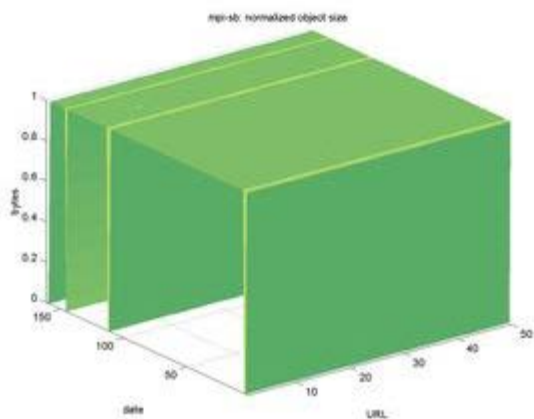
Figure 4. CogPrints Results

3.5 Max-Planck-Institut für Informatik Research Reports (data.mpi-sb.mpg.de)

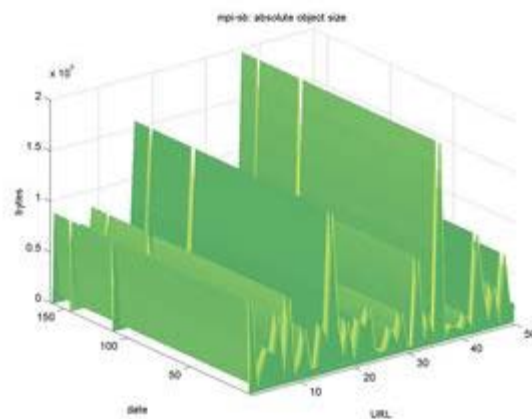
The entry page for Max-Planck-Institut für Informatik (MPI) is <http://data.mpi-sb.mpg.de/reports>, and it is maintained by the MPI library. The DL contains reports authored by MPI staff, and contains a mixture of DVI and PostScript formats. The DL itself is implemented using a Lotus Notes database, and URLs are in the form of:

<http://data.mpi-sb.mpg.de/internet/reports.nsf/c125634c000710d0c12560400034f45a/78>

25bf6e89401969c125607c0062cc62/\$FILE/MPI-I-94-223.dvi Figure 5 shows the results for MPI. Except for five days when the entire DL was inaccessible (2000-11-26, 2000-11-28, 2001-08-12, 2001-08-14, 2001-11-02), all objects were available and consistently delivered exactly the same size as the baseline.



Max-Planck-Institut für Informatik (normalized
byte count)
[MPEG Rotation](#)



Max-Planck-Institut für Informatik (absolute
byte count)
[MPEG Rotation](#)

Figure 5. MPI Results

3.6 IDEAS (ideas.uqam.ca)

The entry page for IDEAS is <<http://ideas.uqam.ca/ideas>>. IDEAS is part of the RePEc system, a set of cooperating archives for research papers in economics. RePEc collects the metadata for author contributed papers and the actual papers remain stored at the local sites. IDEAS is actually one of several search interfaces to the RePEc collection; anyone can set up a search service on the RePEc data. Generally, objects in RePEc are directly accessible as HTML, PostScript or PDF files, such as:

<http://cowles.econ.yale.edu/P/cd/d12b/d1272.pdf>

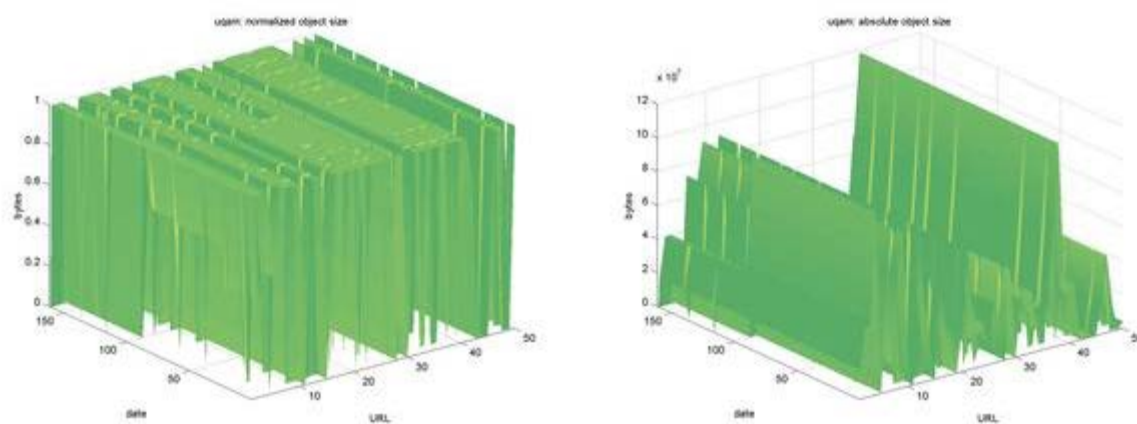
Figure 6 shows the results for IDEAS. As one would expect for a DL with completely distributed object storage, the availability of some objects is sporadic. In particular, there are 8 objects that were initially included, but were never found:

<http://markov.commerce.ubc.ca/marty/patent.rev.ps>
<ftp://ftp.econ.umn.edu/outgoing/geweke/papers/paper63/>
<ftp://ftp.tinbergen.nl/pub/papers/DPs/00055.pdf>
<ftp://weber.ucsd.edu/pub/econlib/dpapers/ucsd9703.ps.gz>
<http://www.csuohio.edu/changm/medusa.pdf>
<http://www.ssc.wisc.edu/~noomes/boston.pdf>
<http://ief.unimc.it/castellano/>
<ftp://ftp.tinbergen.nl/pub/papers/DPs/00029.pdf>

Of those 8, we were able to find three of the objects at new URLs:

ftp://ftp.tinbergen.nl/pub/papers/DPs/00XXX_Series/00055.pdf
ftp://ftp.tinbergen.nl/pub/papers/DPs/00XXX_Series/00029.pdf
<http://www.biz.uiowa.edu/faculty/jgeweke/papers/paper63/>

one of which reflected an institution change by the author, and the other two reflecting a restructuring of the ftp server. The other five objects from IDEAS are presumed to be lost.



IDEAS (normalized byte count)
[MPEG Rotation](#)

IDEAS (absolute byte count)
[MPEG Rotation](#)

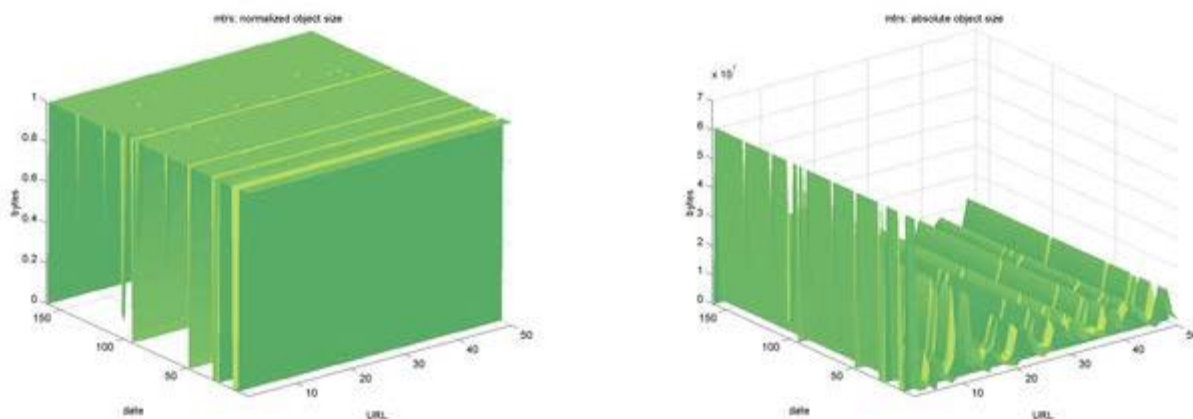
Figure 6 IDEAS Results

3.7 Marshall Technical Report Server (mtrs.msfc.nasa.gov)

The Marshall Technical Report Server (MTRS) contains PDFs of technical reports by NASA Marshall Space Flight Center (MSFC) authors and contractors. MSFC provides engineering support for space flight operations, and conducts research on space propulsion, transportation and microgravity. MTRS is part of the NASA Technical Report Server (NTRS), which is a collection of MTRS-like nodes for all of NASA. URLs are of the form:

<http://mtrs.msfc.nasa.gov/mtrs/2000/cp210429.pdf>

Figure 7 shows the results for MTRS. Though frequently inaccessible (2000-12-10, 2000-12-12, 2000-12-15, 2000-12-19, 2001-01-14, 2001-01-16, 2001-01-26, 2001-03-11, 2001-06-29), MTRS objects remained available and frequently of the same size as the baseline. Some of the larger objects would occasionally transmit fewer bytes than was expected, but this is likely due to network difficulties.



Marshall Technical Report Server (normalized
byte count)

[MPEG Rotation](#)

Marshall Technical Report Server (absolute
byte count)

[MPEG Rotation](#)

Figure 7. MTRS Results

3.8 PubMed (pubmedcentral.nih.gov)

PubMed Central is a digital library of life sciences publications managed by the National Library of Medicine (NLM). NLM is part of the National Institutes of Health (NIH), and administers many digital collections of health sciences information. Initially, PubMed Central had objects with URLs similar to:

<http://www.pubmedcentral.nih.gov/picrender.cgi?artid=14818&picotype=5>

However, as of 2001-07-13 all of these URLs ceased working. This can be seen in Figure 8 as bytes transferred suddenly drop off to zero. Object URLs are now of the form:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=60994>

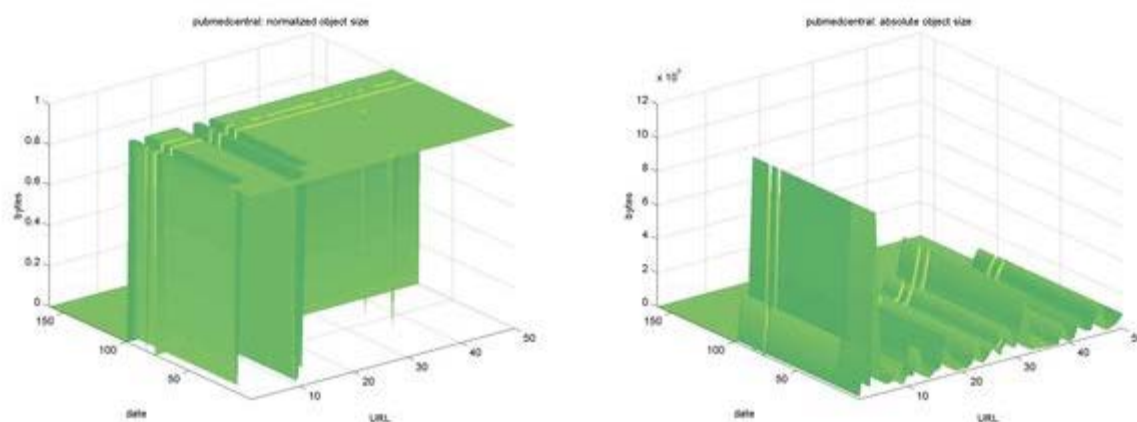
Performing this substitution, we are still unable to find 24 objects:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=15634>

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16071>

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16066>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9233>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9234>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9235>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9239>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9240>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9241>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16065>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16067>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16068>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16069>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16070>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9242>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9243>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9244>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=9245>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=11318>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=11387>

The status of these objects is unknown: they may be temporarily unavailable, removed from the database, or have new article identifiers.



PubMed Central (normalized byte count)
[MPEG Rotation](#)

PubMed Central (absolute byte count)
[MPEG Rotation](#)

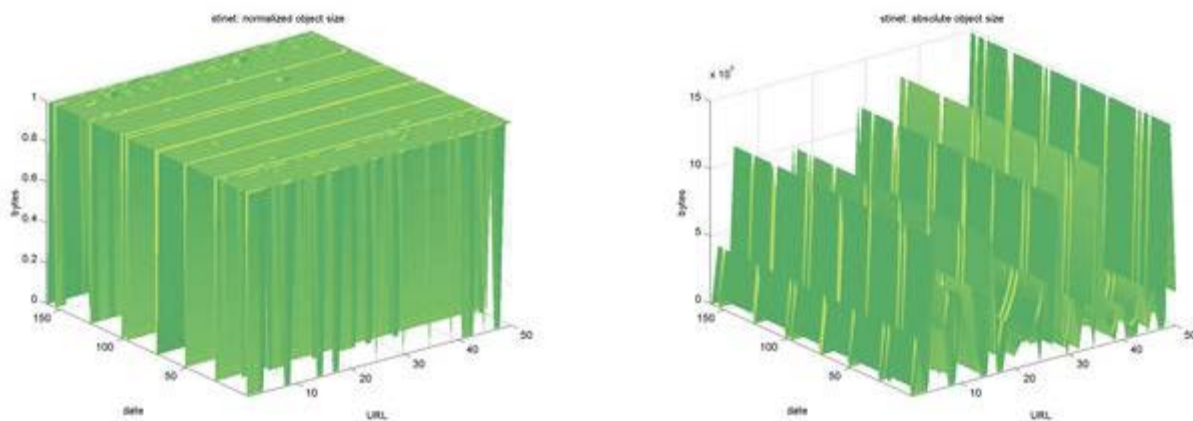
Figure 8. PubMed Central Results

3.9 Scientific and Technical Information Network (STINET) (stinet.dtic.mil)

The Scientific and Technical Information Network (STINET) is run by the Defense Technical Information Center (DTIC). STINET contains information related to the U.S. Government's defense related research. Their primary customers are U.S. Government employees and contractors, but some of the information is publicly available without restriction. We extracted objects from their full text DL available at <<http://stinet.dtic.mil/str>>. The objects are stored in a Fulcrum database and have URLs similar to:

http://stinet.dtic.mil/cgi-bin/fulcrum_main.pl?database=FT_U2&searchid=9730524781058&keyfieldvalue=ADA359339&filename=%2Ffulcrum%2Fdata%2FTR_fulltext%2Fdoc%2FADA359339.pdf

Figure 9 shows the results for STINET. Despite a number of days when the entire DL was unreachable (2001-01-21, 2001-03-18, 2001-05-11, 2001-05-13, 2001-07-08, 2001-07-17, 2001-09-16, 2001-11-23), all of the objects remained available. Some of the objects had small change variations, and a possible explanation is that some of the retrievals of the larger reports timed out while being downloaded.



Scientific and Technical Information Network
(normalized byte count)
[MPEG Rotation](#)

Scientific and Technical Information Network
(absolute byte count)
[MPEG Rotation](#)

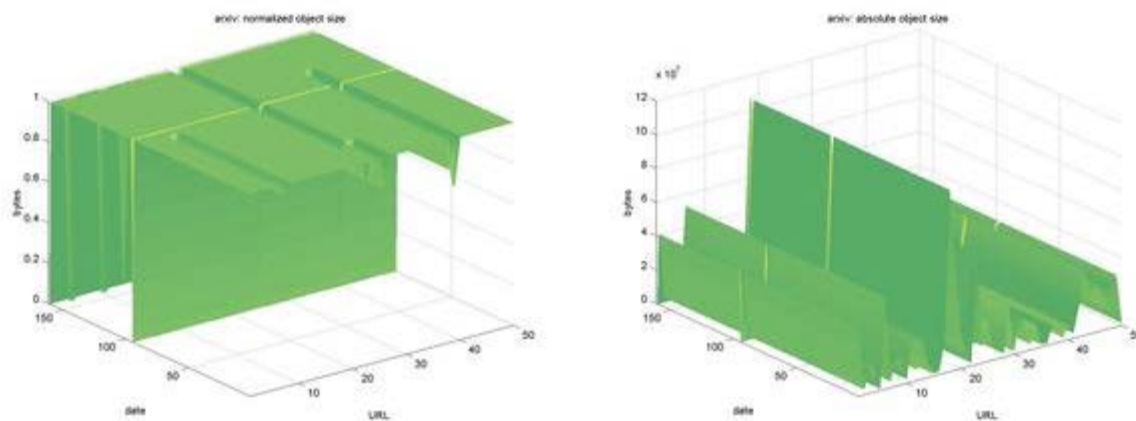
Figure 9. STINET Results

3.10 arXiv.org e-Print archive (www.arxiv.org)

The arXiv.org e-Print archive, formerly known as the LANL e-Print archive, is perhaps the oldest author contributed pre-print DL. It initially focused on high-energy physics, but eventually expanded to cover all of physics, mathematics and computer science. Most of the objects in arXiv are submitted as TeX / LaTeX source files, and other formats such as PostScript and PDF are dynamically generated from them. To avoid the problem of dynamically generating formats on arXiv, where possible we selected the TeX / LaTeX source files in URLs similar to:

<http://www.arxiv.org/e-print/astro-ph/0010324>

When the source files were not available, we choose PDF formats. Figure 10 shows the results for arXiv. There was only one day in which the entire DL was unreachable (2001-07-01). The change in byte size for some of the objects is due to the fact that arXiv allows revisions to be posted for an object, while keeping the same accession number. All objects in arXiv remained available.



arXiv e-Print archive (normalized byte count)

[MPEG Rotation](#)

arXiv e-Print archive (absolute byte count)

[MPEG Rotation](#)

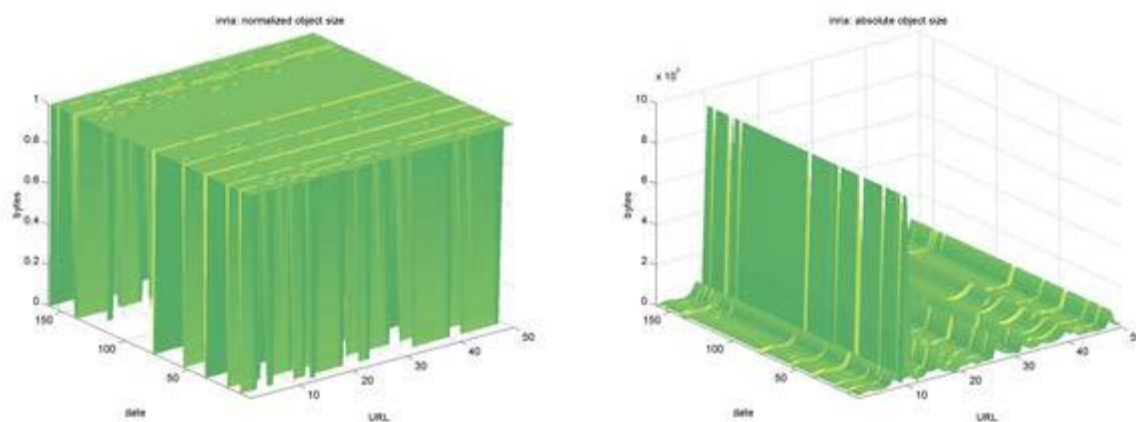
Figure 10. arXiv Results.

3.11 Research Reports, Technical Reports, Theses edited by INRIA (www.inria.fr)

INRIA is the French national research institute for computer science. Their English language interface to their DL is located at <http://www.inria.fr/rrrt/index.en.html>, and it contains reports, theses and e-prints authored by INRIA staff. Objects are available in PostScript and PDF formats and have URLs similar to:

<ftp://ftp.inria.fr/INRIA/publication/RT/RT-0188.ps.gz>

Figure 11 shows the results for INRIA. Although the entire DL was unreachable for only two days (2001-02-11, 2001-02-13), there were many days when network difficulties prevented access to a great number of the objects. However, all of the objects remained available.



INRIA (normalized byte count)

[MPEG Rotation](#)

INRIA (absolute byte count)

[MPEG Rotation](#)

Figure 11. INRIA Results

3.12 Jefferson Lab Research Publications (www.jlab.org)

The Thomas Jefferson National Accelerator Facility (Jefferson Lab, or JLab) is a nuclear physics laboratory, formerly known as the Continuous Electron Beam Accelerator Facility. A list of pre-prints by JLab staff is available from http://www.jlab.org/div_dept/admin/publications. The objects are available in PDF format with URLs similar to:

http://www.jlab.org/div_dept/admin/publications/papers/00/PHY00-15.pdf

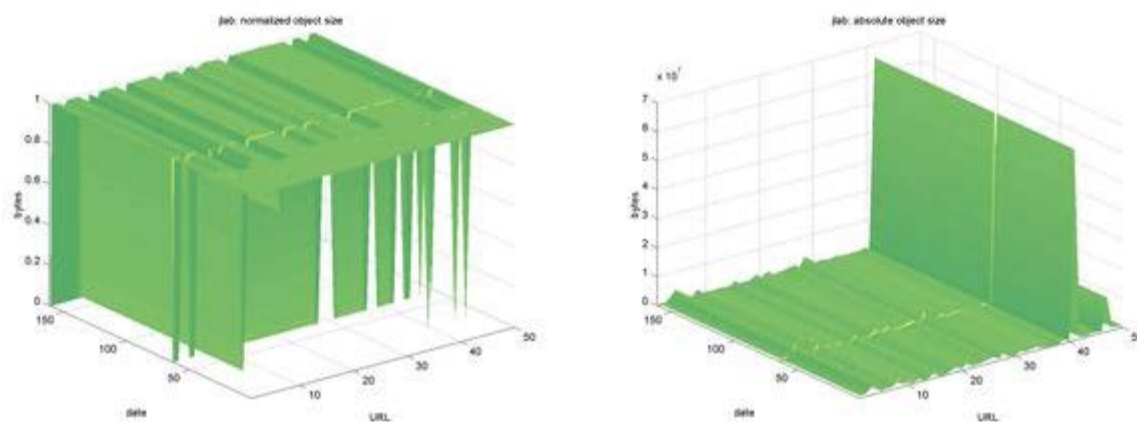
Figure 12 shows the results for JLab. One of the objects at first appears to have been lost:

http://www.jlab.org/div_dept/admin/publications/papers/00/ACP00-02.PDF

However, further investigation reveals that the URL has had its capitalization removed, and the object is now available at:

http://www.jlab.org/div_dept/admin/publications/papers/00/ACP00-02.pdf

A small number of the objects appear to have slightly shrunk after the first half of the testing period (they appear as shallow troughs in the normalized graph in Figure 12). Since the JLab PDFs appear to be generated from scanned pages, perhaps the PDF was regenerated with better compression from the original TIFF files. All JLab objects remain available.



Jefferson Laboratory (normalized byte count) [MPEG Rotation](#) Jefferson Laboratory (absolute byte count) [MPEG Rotation](#)

Figure 12. JLab Results

3.13 Networked Computer Science Technical Reference Library (www.ncstrl.org)

The Networked Computer Science Technical Reference Library (NCSTRL) is a DL with over 100 international participants. NCSTRL contains technical reports and e-prints in computer science and related disciplines. Originally based on the Dienst software package and protocol ([Davis & Lagoze, 2000](#)), the original NCSTRL ceased operation in October 2001 (approximately the last 1/3 of the time period tested). During this time frame, NCSTRL was transitioned to an Open Archives

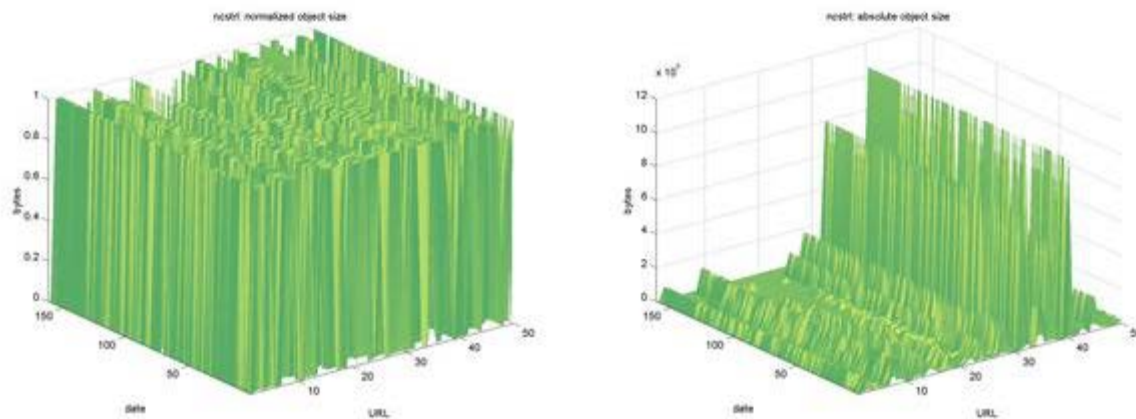
Initiative (OAI) based implementation ([Lagoze & Van de Sompel, 2001](#)). However, some individual Dienst servers at various institutions continued to function. Objects in the OAI-based implementation were not tested; only Dienst objects were tested, and they had URLs similar to:

http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.uwash_cse/TR-95-11-05?format=postscript

Figure 13 shows the results for NCSTR. The responses are highly variable and give some insight into the decision to end the Dienst-based implementation. There appear to be 38 objects missing from the Dienst-based NCSTR:

<http://xxx.lanl.gov:80/Dienst/Repository/1.0/Disseminate/xxx.cs.DC:9809019?part=page=&content-type=application/postscript>
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.cabernet/BROADCAST#TR95-82>
http://cstr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ercim.inria.publications/RR-2595?format=pdf_ftp_1
http://cstr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ercim.inria.publications/RR-3147?format=pdf_ftp_1
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.mit_lcs/MIT/LCS/TM-384
http://cstr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ercim.inria.publications/RR-2767?format=pdf_ftp_1
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.ucb/CSD-87-334?format=ocr>
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.tucs.fi/TUCS-TR-289>
<http://cstr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.cabernet/BROADCAST#TR94-38>
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.uci/ICS-TR-95-19>
http://cstr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ercim.inria.publications/RR-2394?format=pdf_ftp_1
<http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/GetPage/ncstrl.cmu/CS-93-231?format=postscript>
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.ucb/CSD-86-268>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.ethz_cs/1994TR-213
<http://cstr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.cabernet/BOLOGNA#UBLCS-92-1>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.uwash_cse/TR-98-10-03?format=postscript
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.uci/ICS-TR-92-26>
<http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/GetPage/xxx.cs.DC/9809019?format=application/pdf>
http://cstr.cs.cornell.edu:80/Dienst/UI/1.0/GetPage/ercim.inria.publications/RR-3119?format=pdf_ftp_1
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.ncsu_cs/TR-96-18
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ercim.forth.ics/TR100-0266>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.tamu_cs/TR95-009
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.ustuttgart_fi/TR-1999-06?format=pdf
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.ustuttgart_fi/TR-1994-13?format=pdf
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.uwash_cse/TR-98-09-01?format=postscript
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.ucsc_cse/UCSC-CRL-90-63?format=postscript
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.rice_cs/TR97-276
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.ucb/CSD-88-422?format=ocr>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.ucla_cs/970044
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/caltechCSTR/1997.cs-tr-97-07?format=postscript>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.odu_cs/TR_97_26?format=postscript
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.uci/ICS-TR-96-20>
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.cabernet/NEWCASTLE-CS#TR94-499>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.auburn_eng/CSE94-14?format=postscript
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.icas/TR-2000-4?format=postscript>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.uwash_cse/TR-95-11-05?format=postscript
<http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Download/ncstrl.umcp/CS-TR-4003>
http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/GetPage/ncstrl.odu_cs/TR_93_27?format=postscript

We were able to find all of these 38 objects in other areas: either the new NCSTR, other DLs that had copied NCSTR content, or from the original departmental servers themselves. Since it actually ceased operations and transferred to new implementation during our testing, NCSTR is a special case in our experiment. For the purposes of this study, we will consider that no objects were lost.



NCSTRL (normalized byte count)

[MPEG Rotation](#)

NCSTRL (absolute byte count)

[MPEG Rotation](#)

Figure 13. NCSTRL Results

3.14 New Zealand Digital Library (www.nzdl.org)

The New Zealand Digital Library (NZDL) is a multidisciplinary DL with many collections run by the University of Waikato. For the purposes of this study, we examined only the computer science technical report section of NZDL. NZDL maintains an index of computer science reports held at various institutions. The objects we selected were PostScript, and had URLs similar to:

```
http://www.nzdl.org/fast-cgi-bin/cstrlibrary?e=d-0cstr--00-1--1011--1-en-50---20
-about-distributed+computing--21-00-0w
-0&cl=search&d=HASH01037986be25db2f92db08dd.1&gg=postscript&gga=ftp%3a
%2f%2fagora%2eleads%2eac%2euk%2fscs%2fdoc%2freports%2f1994%2f9
4%5f17%2eps%2eZ
```

Figure 14 shows the results for NZDL. There were several days in which the entire DL was not available (2001-04-03, 2001-04-20, 2001-04-29, 2001-04-01, 2001-05-11, 2001-05-13, 2001-05-15, 2001-05-18, 2001-09-07, 2001-10-21, 2001-10-28, 2001-11-30), as well as a handful of individual download errors. There are two objects: that appear unavailable:

```
http://www.nzdl.org/fast-cgi-bin/cstrlibrary?e=d-0cstr--00-1--1011--1-en-50---20
-about-distributed+computing--21-00-0w-0
&cl=search&d=HASH42dd0432278d6ece6ab6ac.1&gg=postscript&gga=ftp%3a
%2f%2factor%2ecs%2evt%2eedu%2fpub%2fact++%2fdoc%2fuser%5fman%2eps
```

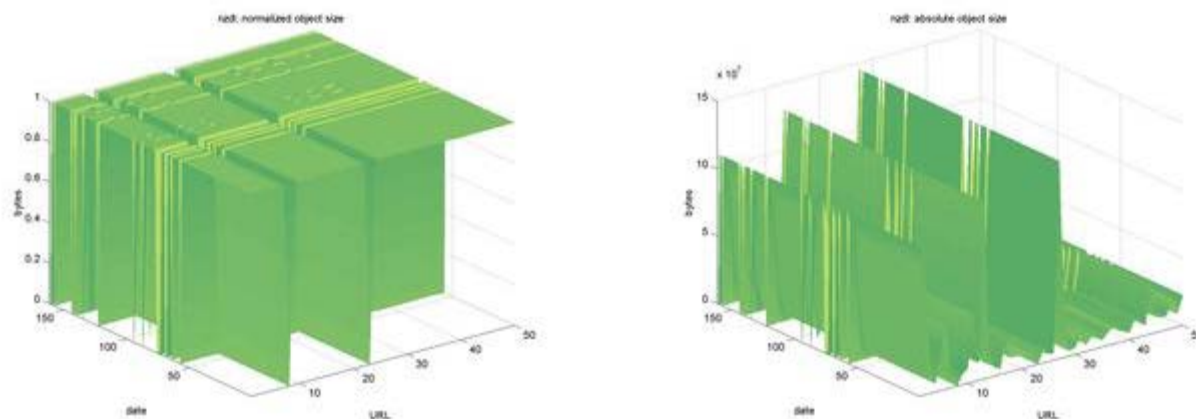
```
http://www.nzdl.org/fast-cgi-bin/cstrlibrary?e=d-0cstr--00-1--1011--1-en-50---20
-about-distributed+computing--41-00-0w
-0&cl=search&d=HASHf9cc26931ee3fc77da7644.1&gg=postscript&gga=ftp
%3a%2f%2factor%2ecs%2evt%2eedu%2fpub%2fact++%2fdoc%2fjoop%2eps
```

These appear to be missing because they point to a machine, "actor.cs.vt.edu", that no longer has a DNS entry. The NZDL has cached copies of these (in GIF format, extracted from the original PostScript) available at the following respective URLs:

```
http://www.nzdl.org/fast-cgi-bin/cstrlibrary?e=q-0cstr--00-1--1011--1-en-50---20
-about-distributed+computing--21-00-0w-
0&a=d&c=cstr&cl=search&d=HASH42dd0432278d6ece6ab6ac
```

<http://www.nzdl.org/fast-cgi-bin/cstrlibrary?e=d-0cstr--00-1--1011--1-en-50---20-about-distributed+computing--41-00-0w-0&cl=search&d=HASHf9cc26931ee3fc77da7644.1>

So while the original PostScript copies of these documents are no longer available from the original site, the NZDL cached copies continue to be available. Because of this, we consider all objects in the NZDL to be available.



New Zealand Digital Library (normalized byte count)

[MPEG Rotation](#)

New Zealand Digital Library (absolute byte count)

[MPEG Rotation](#)

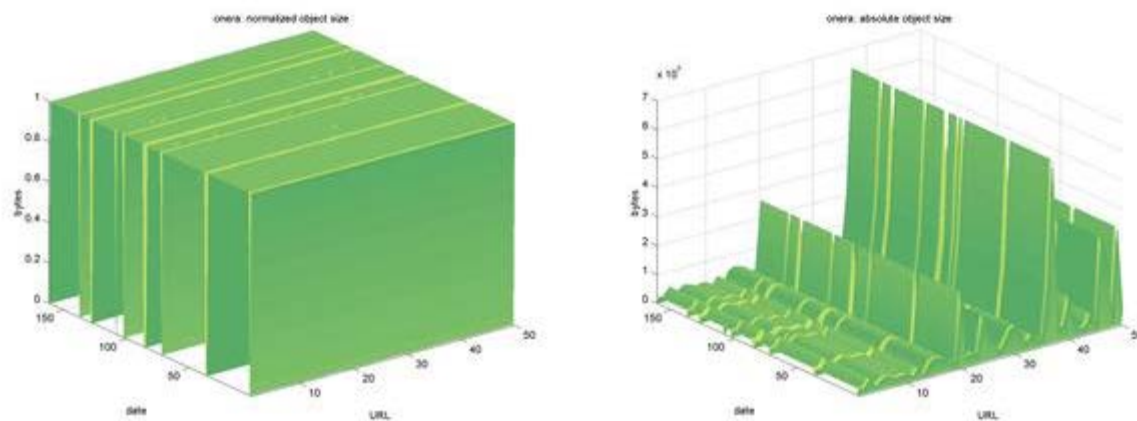
Figure 14. NZDL Results

3.15 ONERA Publications (www.onera.fr)

ONERA is the French national aerospace research institute. The English language interface for their DL is available at http://www.onera.fr/SEARCH/BASIS/public/web_en/document/sf. It contains reports, pre-prints and e-prints published by ONERA staff members. Objects are available in PDF format, stored in a BASIS database and have URLs similar to:

http://www.onera.fr/SEARCH/BASIS/public/web_en/document/DDD/302166.pdf

Figure 15 shows the results for ONERA. Other than several days in which the entire DL was not available (2000-11-24, 2001-02-11, 2002-02-13, 2001-05-08, 2001-06-08, 2001-06-10, 2001-07-20, 2001-09-21, 2001-09-23, 2001-10-14), the objects remained available.



ONERA (normalized byte count)
[MPEG Rotation](#)

ONERA (absolute byte count)
[MPEG Rotation](#)

Figure 15. ONERA Results

3.16 RAND's Publication Database (www.rand.org)

RAND is a multidisciplinary research and public policy institute, first created by the U.S. military in 1946. The DL of unrestricted publications by their staff is available at <http://www.rand.org/Abstracts>. Their objects are available in PDF format with URLs similar to:

<http://www.rand.org/publications/MR/MR650.pdf>

Figure 16 shows the results for RAND. There were two days in which the entire DL was unavailable (2001-06-01, 2001-07-22). The normalized byte count graph shows a lot of variability. At first glance, there appears to be missing to be missing objects. However, it turns out that many of the RAND objects have been restructured during the testing period. Some of them grew in size, and two of the objects that were previously available at a single URL:

<http://www.rand.org/publications/MR/MR994/MR994.pdf>

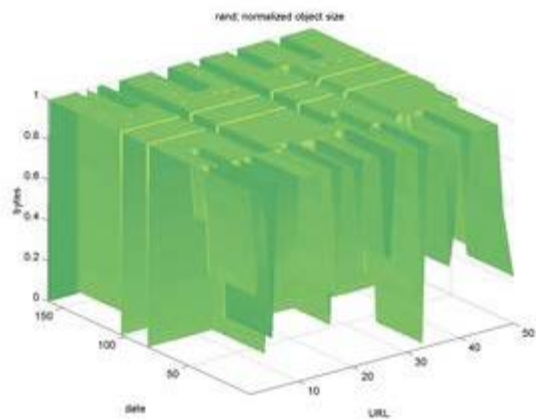
<http://www.rand.org/publications/MR/MR614/MR614.pdf>

Have now been turned into multiple chapters, with a single PDF file for each chapter. The correct URLs for these two objects are now:

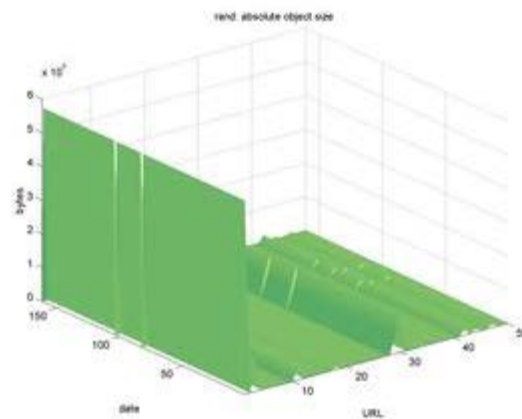
<http://www.rand.org/publications/MR/MR994/>

<http://www.rand.org/publications/MR/MR614/>

Therefore although the objects in the RAND DL underwent significant change during the testing period, they remained available.



RAND (normalized byte count)
[MPEG Rotation](#)



RAND (absolute byte count)
[MPEG Rotation](#)

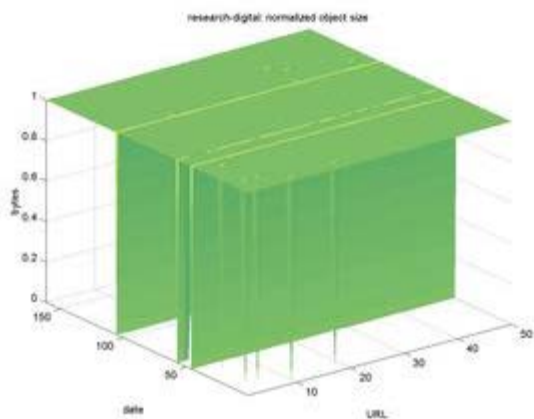
Figure 16. RAND Results

3.17 COMPAQ Western Research Lab Publications (www.research.digital.com)

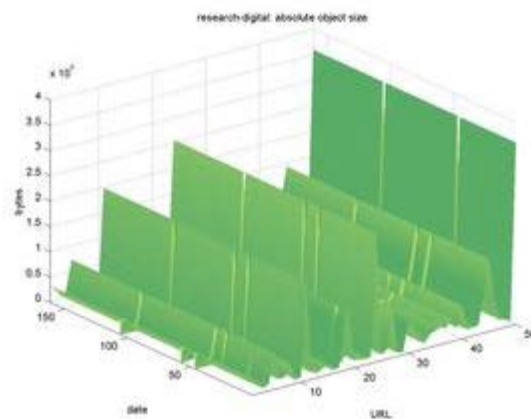
The COMPAQ Western Research Laboratory (formerly Digital Equipment Corporation's Western Research Laboratory) contains publicly available technical reports by COMPAQ (and former DEC) research staff. This DL can be found at <http://www.research.compaq.com/wrl/admin/publications.html>, and it offers PostScript and PDF versions of objects with URLs similar to:

<ftp://gatekeeper.dec.com/pub/DEC/WRL/research-reports/WRL-TR-93.1.pdf>

Figure 17 shows the results for COMPAQ. Except for two days when the entire DL was unavailable (2001-03-06, 2001-07-27) and a handful of failed individual downloads, the objects in the COMPAQ remained available.



COMPAQ Western Research Lab (normalized byte count)
[MPEG Rotation](#)



COMPAQ Western Research Lab (absolute byte count)
[MPEG Rotation](#)

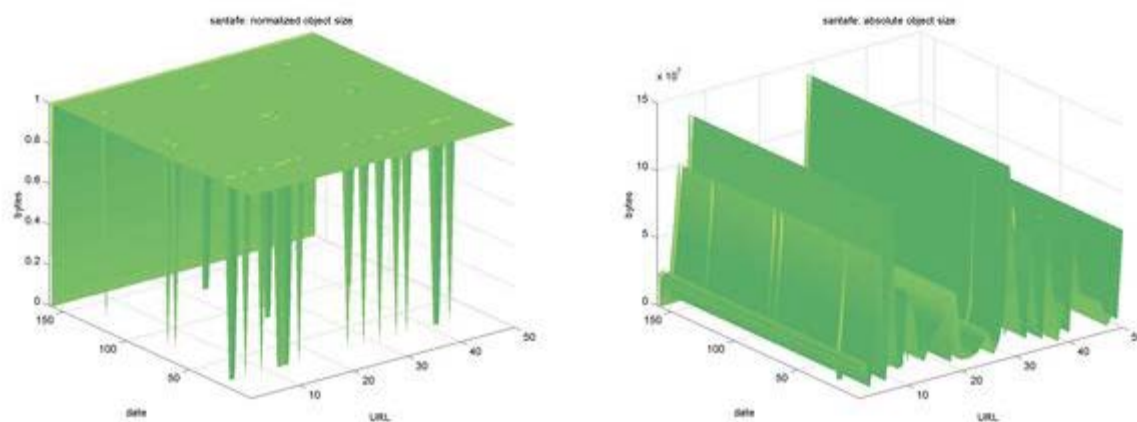
Figure 17. COMPAQ Results

3.18 Santa Fe Institute Publications (www.santafe.edu)

The Santa Fe Institute (SFI) is a private, multidisciplinary, visiting research institution that has historical ties to the Los Alamos National Laboratory. The SFI DL is at <http://www.santafe.edu/sfi/indexPublications.html> and holds working papers by the SFI staff and visiting researchers. It has objects in PostScript and PDF formats, with URLs similar to:

<http://www.santafe.edu/sfi/publications/Working-Papers/00-11-061.ps.gz>

Figure 18 shows the results for SFI. Except for a single day when the entire DL was unavailable (2001-12-04) and a few isolated failed downloads, the objects in the SFI DL remained available.



Santa Fe Institute (normalized byte count)

[MPEG Rotation](#)

Santa Fe Institute (absolute byte count)

[MPEG Rotation](#)

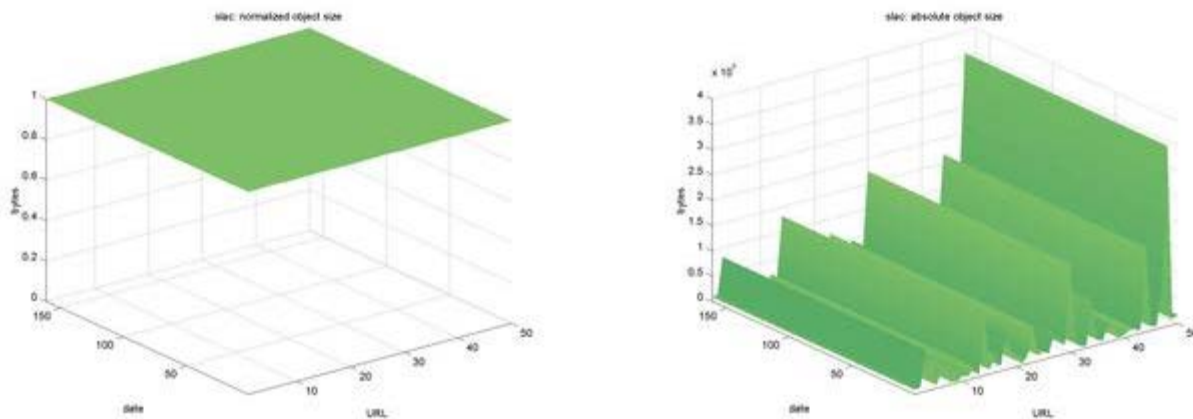
Figure 18. SFI Results

3.19 SLAC Document Server (www.slac.stanford.edu)

The Stanford Linear Accelerator Center (SLAC) is a high energy physics research facility located on the campus of Stanford University. The DL for SLAC can be accessed at <http://www.slac.stanford.edu/pubs/fastfind.html>. It contains physics pre-prints from many other laboratories, and its objects are available in PDF and PostScript format with URLs similar to:

<http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-8688.pdf>

Figure 19 shows the results for SLAC. The SLAC DL is unique in that during testing it did not experience a single download failure. The results from testing the SLAC DL represent the best possible object availability.



Stanford Linear Accelerator Center (normalized
byte count)

[MPEG Rotation](#)

Stanford Linear Accelerator Center (absolute
byte count)

[MPEG Rotation](#)

Figure 19. SLAC Results

3.20 Electronic Theses and Dissertations (www.theses.org)

The Electronic Theses and Dissertations (ETD) project is a broad project including over 100 institutions. However, for the purposes of this project, we included only objects culled from the Virginia Tech portion of the collection, the home page of which can be accessed at <http://scholar.lib.vt.edu/theses>. Some of the objects in the ETD are restricted access; available only to those at the home institution. For this study, we chose only unrestricted objects. The ETD offers objects in PDF format with URLs similar to:

http://scholar.lib.vt.edu/theses/available/etd-08312000-16220053/unrestricted/Thesis_final.pdf

Figure 20 shows the results for ETD. There was one day in which the entire DL was unreachable (2001-11-04). There are two interesting objects in the results. There is one object that appears missing:

<http://scholar.lib.vt.edu/theses/available/etd-51298-1519/unrestricted/phdthesis.pdf>

Even the root URL of:

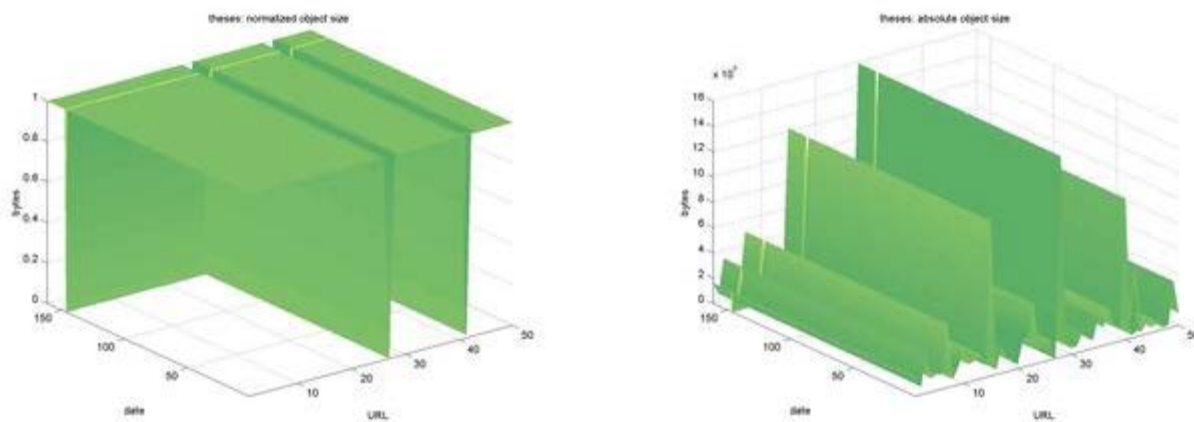
<http://scholar.lib.vt.edu/theses/available/etd-51298-1519/>

is not available. Perhaps this was never really an object, and this URL is a result of a typographical or administrative error. We consider this to be a lost object.

There is a second interesting object:

<http://scholar.lib.vt.edu/theses/available/etd-09102000-22560027/>

that saw its size drop from 5 megabytes to 10 kilobytes. This probably represents a change in the objects structure, going from the default of yielding the PDF file to yielding an HTML page describing the PDF file (which remains available). We do not consider this to be a lost object, but rather simply one whose structure was brought into alignment with the other objects in the DL shortly after testing began.



Electronic Theses and Dissertations (normalized
byte count)

[MPEG Rotation](#)

Electronic Theses and Dissertations (absolute
byte count)

[MPEG Rotation](#)

Figure 20. ETD Results

4.0 Conclusions

We expect that objects placed in a DL should persist longer than an average web page; the very act of placing an object in a DL would seem to indicate some intention of long-term preservation. While it is possible that not all objects are deserving of long-term preservation, who is to make this judgment? For example, several of the early Internet RFCs have been lost ([Arms, 1999](#)), and while the working of the Internet has not been compromised by their loss, historians would surely prefer an intact collection.

Further study of the nature of DL object change and persistence is needed. For periods when the entire DL was inaccessible, we can assume that either the host machine was down or network conditions were unfavorable at the time of testing. However, the nature of object unavailability needs to be studied further to determine if the failure lay with the DL itself, or simply artifacts of transient network/http problems.

Furthermore, no attempt was made to determine the "correctness" of the objects during the baseline testing. This, along with investigating the object when byte sizes change, would further reveal the nature of the frequent changes observed. Despite the large number of changes observed in object size and periods of transient unavailability, it appears that 31 of the original 1000 objects are no longer available, or have at least evaded our best attempts to locate them. This matches the value of 3% lost URLs reported by Lawrence et al. (2001). However, further study is needed to determine if 3% is a generalizable number.

References

Arms, W. A. (1999). Preservation of scientific serials: three current examples. *Journal of Electronic Publishing*, 5(2). Available at <<http://www.press.umich.edu/jep/05-02/arms.html>>.

Davis, J. R. & Lagoze, C. (2000). NCSTRL: design and deployment of a globally distributed digital

library. *Journal of the American Society for Information Science*, 51(3), 273-280.

Douglis, F. Feldmann, A. Krishnamurthy, B. and Mogul, J. (1997) Rate of change and other metrics: A live study of the World Wide Web. In *Proceedings of the USENIX on Internet technologies and systems* (pp. 147-158). Available at: http://www.usenix.org/publications/library/proceedings/usits97/douglis_rate.html.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2), 162-180.

Koehler, W. (2000) Digital libraries and World Wide Web sites and page persistence. *Information Research*, 4(3). Available at: <http://InformationR.net/ir/4-4/paper60.html>.

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the first ACM/IEEE Joint Conference on Digital Libraries* (pp. 54-62). Available at: <http://www.cs.cornell.edu/lagoze/papers/oai-jcdl.pdf>.





Lawrence, S., Coetzee, F., Glover, E., Pennock, D., Flake, G., Nielsen, F., Krovetz, R., Kruger, A., and Giles, C. L. (2001). Persistence of web references in scientific research, *IEEE Computer*, 34(2), 26-31. Available at: <http://www.neci.nec.com/~lawrence/papers/persistence-computer01/>.

Powell, A. L. and French, J. C. (2000). Growth and server availability of the NCSTRL digital library. In *Proceedings of the fifth ACM conference on digital libraries* (pp. 264-265). Available at: <http://www.cs.virginia.edu/~cyberia/papers/DL00.pdf>.

Viles, C. L. and French, J. C. (1995). Availability and latency of World Wide Web information servers. *Computing Systems*, 8(1), 61-91.

Appendix

A [compressed tar file](#) (3,340 files, 13.3 MB) is provided with the following structure:

- analyze-data.pl - creates the data.analyzed file
- check.pl - tests the DL objects
- cron.run - the crontab entry used to run check.pl
- dls - the list of DL entry pages used to find the objects
- matlab/ - all the Matlab scripts used to generate the figures and MPEGs
- no-finish - the dates that did not complete running on ruby.ils.unc.edu
- reducer.pl - creates the data files
- test/ - the directory with the test data
 - 1 directory per DL

- 1 file per date tested
- data - the resulting matrix
- data.analyzed - a summary of the data file
- baseline - the initial size of the objects
- objects - the metadata and object URLs

[Top](#) | [Contents](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Back Issues](#)
[Previous article](#) | [Next article](#)
[Home](#) | [E-mail the Editor](#)

[D-Lib Magazine Access Terms and Conditions](#)

DOI: [10.1045/january2002-nelson](#)