Link Accessibility in Electronic Journal Articles

Donna Bergmark*
Cornell Digital Library Research Group

March 31, 2000

1 Introduction

The prevalence of online scholarly, technical work is increasing at a great rate. In 1995, Hitchcock, Carr, and Hall [6] reported that there were 115 scholarly peer-reviewed online journals in Science and Technology, plus a comparable number in the Social Sciences; in 1999 Maclennan [9] estimated that 10,000 electronic journals were in existence. With this increase in electronic literature, the idea of interlinking online documents is growing more attractive.

Cornell is one of several places engaged in exploring methods for reference linking the online literature, partly through its involvement in the Open Citation project,[1] and partly though research being done for CNRI. The DOI-X project is a similar effort being carried out by commercial journal publishers [7].

References are provided by human authors who know the literature; increasingly, these are accompanied by URLs provided by the authors. The reference linking project aims to amplify this effort by looking up online locations for references that are not accompanied by URLs. In either case, the URLs will be followed to access the cited document directly. Reference links are the best tool for repository interoperability.

*CNRI Grant #2057/57-02 and NSF Grant # IIS-9907892

[1] `http://opcit.eprints.org/`

For reference linking to succeed, however, it is important that the links be stable over time. This has led to an effort to define persistent reference links [10]. It is unlikely, however, that the DOI system will extend to all online literature. Davidson [3] in fact argues that not “too many borderline institutions” should “participate in the DOI system.” This would probably include the very popular open archives as being promoted by the Open Archives Initiative [14]. Thus persistence of “DOI-less” links is an issue.

In fact, the whole DOI initiative is based on the assumption that URLS change frequently, thus requiring the invention of persistent identifiers. To estimate this need, we chose to examine the longevity of external links in the D-Lib Magazine, an electronic journal published by the CNRI.

2 Related Work

Other reference link studies have been performed. Five years ago, Harter and Kim [4] analyzed roughly 200 online journals and found that in 1995 very few electronic journal articles cited other online articles, and that about half of those references were inaccessible later in the same year. Thus their work seems to be a grim prognosis for achieving reference-linked online literature, both because the links don’t exist, and those that do become corrupted in short order.

Some work has been done on preservation of content and links. For example, [2] does a case study of three online journals and concludes that preservation of content is easier than preservation of access or preservation of appearance. But if content also includes the references and the associated hyperlinks, it might not be so easy to preserve content since the breaking of links could be beyond the publisher’s control. External links in the D-Lib magazine, for example, are not monitored[2].

3 Reference Linking Strategies

Reference linking refers to turning references within an online document into “live references”; that is, while viewing a paper on your screen, you can follow references in that paper to other network accessible papers and view those as well (in separate windows on your terminal). As stated above, it is especially attractive to develop this technology because of the increasing number of technical journals and magazines available online.

Most commonly, references are found in a late section of an article; this section is often labeled *References*, *Bibliography*, or *List of References*. This makes it relatively easy for a computer program to locate the references of an electronic article. Increasingly these notations include URL's pointing directly to online information. In many cases, references that do not include URLs mention published journal articles that can also be resolved to an online copy. For example, ACM journals can be found in print as well as online in the ACM online library, assuming that the reader or the reader's institution is a subscriber to the ACM digital library.

The two key parts of reference linking are: link resolution and display of reference linked full text. Link resolution can include parsing the reference to determine what is being cited, finding online locations for that paper, and then retrieving the cited paper. If the reference already contains a URL, the reference link is already resolved; the only task is to retrieve the referenced URL.

Link resolution can be done statically or dynamically. The ResearchIndex [8] and Open Journal [5] projects do static link resolution, building citation databases with canonical representation of sources and targets. This is most efficient, because the databases can be updated overnight, online retrieval does not have to wait for all of link resolution to occur, and each new paper need have its references processed only once.

The alternative is to resolve the links on the fly, as is done in the SFX [11, 12, 13] system. This has the advantage that if a paper has moved its location, it can still be retrieved. In either case, the product will be more or less helpful to the researcher depending on the integrity of the links. Reference links are useless if the page being referred to has disappeared or moved.

Before proceeding with a full-scale reference linking project, we were curious to see how much of a problem link decay might be. As stated above, the D-Lib[2] Magazine is of particular interest. This journal exists only on-line and has been putting out monthly issues since mid-1995. A number of static reference links were inserted by authors of D-Lib articles as the article was written.

[2]The magazine can be found at `http://www.dlib.org/dlib`

4 Experimental Approach

The D-Lib Magazine exists as a series of HTML files. It is thus relatively straightforward to examine its external links to see which are broken. The first step was to locate a web-crawler, or robot, that would do this analysis. A candidate was found in the CPAN library[1], WWW::Robot. This is a perl5 module that makes use of existing perl5 networking modules HTML, HTTP, URI, and WWW, as well as wwwlib-perl.

The WWW::Robot module takes a number of parameters, such as a starting URL, the depth to which links should be followed, and whether the tree should be searched depth-first or breadth-first.

It also obeys the standard guidelines for robots, in that it examines `robot.txt` files for permission to follow links. If no such file exists, or if permission is granted, then this link is a candidate for being followed, as long as it is within the specified depth from the starting URL. The link is queued up for examination, either at the head of the queue (breadth-first search) or at the end of the queue (depth-first search). No link is followed more than once, even if it occurs several times during the walk.

To use the robot, one writes a separate perl script that creates a Robot object, and then registers certain handlers with the robot, such as code to execute when a new link is encountered, or when page content is retrieved, or whether to follow the new link.

If one keeps track of which URLs refer to other ones, statistics can be kept on how many links there are on a page, how many of them are new (not yet seen in this run), and how many of the new links are broken. The program developed for this project builds a shadow tree that mimics the node walk behind done by the Robot.

For this project, handlers were provided for the following events:

invoke-on-all-url This handler is called only once per page. It adds a new node to the shadow tree, and initializes its link counters (total, new, and broken) to zero. The number of distinct links for the parent[3] of this page is incremented. The new node's depth in the tree is one more than the parent's.

follow-url-test Here links that should not be followed, such as `http://www.dlib.org/dlib.html` and `http://www.hpcc.gov/`, are weeded out.

[3] A node's parent is the page containing a link to that node.

These sites and pages are not germane to the link analysis of D-Lib articles. In addition, if the URL is at the bottom of the tree, it is simply pinged (its HEAD is gotten) rather than followed.

invoke-on-get-error Add one to the number of broken links on the parent's page, and note the error code (403, 404, 500...).

invoke-on-link A link has been encountered on the current page. Add one to the counter of links on this page.

generate-report This handler is invoked when the Robot's queue of URLs becomes empty. A final report is printed with the statistics that have been accumulated. The statistics included for each followed URL include the number of links on that page, the number of distinct links on that page, the number of distinct links that were broken, and the URL itself. In addition each URL, whether pinged or followed, is annotated either with an OK or the error code that was returned by GET.

Using this robot, two online journals were examined: D-Lib Magazine and the Journal of Electronic Publishing (as a cross-check). Both started in 1995. The D-Lib Magazine is about Digital Libraries, focusing on topics such as persistent identifiers, digital archives, reference linking and the like. The Journal of Electronic Publishing initially paid a lot of attention to financing high speed networks and scholarly publishing on the same, but now has a wide variety of articles relating to electronic publishing, including some reprints.

The D-Lib Magazine has about five articles per issue; the issues come out monthly except for a combined July/August issue every year. Reference links contained in articles and book reviews were studied.

JEP comes out quarterly at this point, though it started out on a less frequent basis than that. It has about ten articles per issue. Reference links contained in the articles and preprints were examined.

In both cases, references are both internal (i.e. to other issues of the magazine) and external (could go anywhere). Notes, book reviews, archived papers are included in the analysis if they seemed to be feature length, with references.

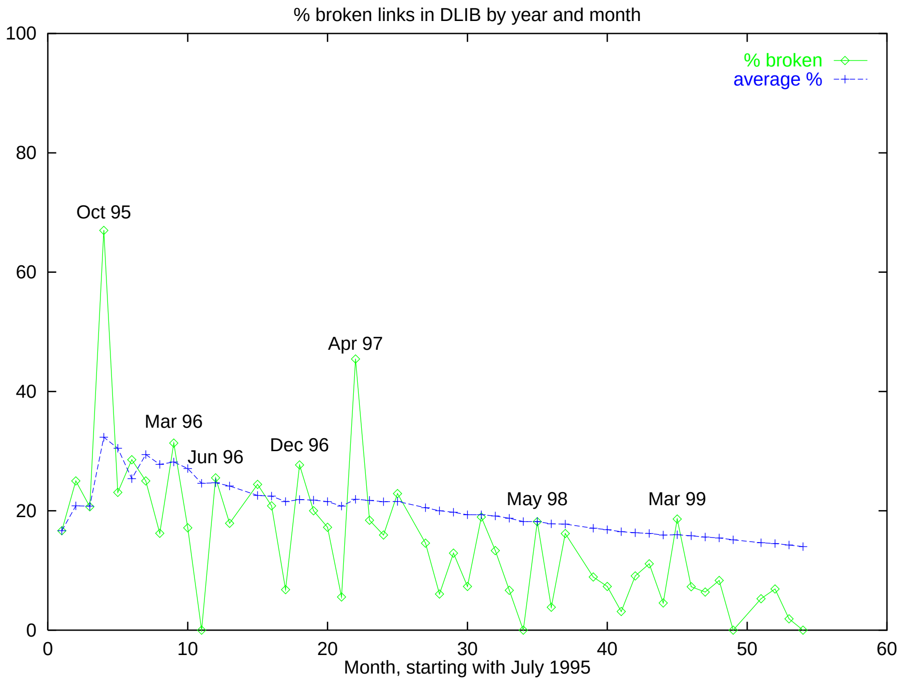

Figure 1: Links Broken and Cumulative Average

5 Results

We were not quite sure what to expect of our analysis. We suspected that the oldest links (from mid-1995) might be as much as 50 % broken, whereas the most recent ones should all work.

The 50% estimate comes in part from the Harter and Kim [4] study of electronic journals. From a fairly large sample, Harter and Kim found that 279 selected on-line articles from electronic journals cited 83 other on-line sources. This represented only 1.9% of the total references (counting only the first 20 references per article). Our results show that the D-Lib magazine is far richer in reference links to online literature: about 10 links per article.

In 1995, Harter and Kim found that only about half of the referenced online resources were actually accessible online. But they include web pages, email messages, newsgroup postings. Today on-line references are far more likely to be HTTP links, which were barely used in 1995. Today, five of their categories would now probably be expressed as URLs: Web Page, Electronic Personal Paper, E-journal Article, Electronic Preprint, and Electronic Newspaper. Considering only these categories (excluding E-mail messages, and the like) Harter and Kim's results showed that in 1995, 72.2% of these were accessible in the same year they were published. Thus their grim results were partially colored by the greater usage of "junk links" in 1995 than now.

In fact, things have gotten even better than that. Figure 1 shows 4.5 years worth of D-Lib issues, with the per cent of links that were broken per issue (considering only articles and book reviews). In very few issues did the per cent broken reach above 30%, and as expected the more recent articles had fewer broken links. Four issues had no broken links at all.

The line graphed with "+" is the average of all the points to the left. The % broken links per issue is approaching an overall level of 14% as of December 1999. It will be interesting to see how this graph changes in a longitudinal study. That is, if the same experiment were run in June 2000, what would the curves look like then?

Plotting the average over all issues gives each successive issue smaller impact on the measured result; thus the initial issues have an excessive influence on the evolving average. To see how the link quality changes over time, a 5-point moving average was plotted. See Figure 2 for this graph. The average % of broken links in articles in the first five issues is around 31%, whereas the average for the last five is only 3%.

Given the funds for this research (from CNRI), the main interest is D-Lib

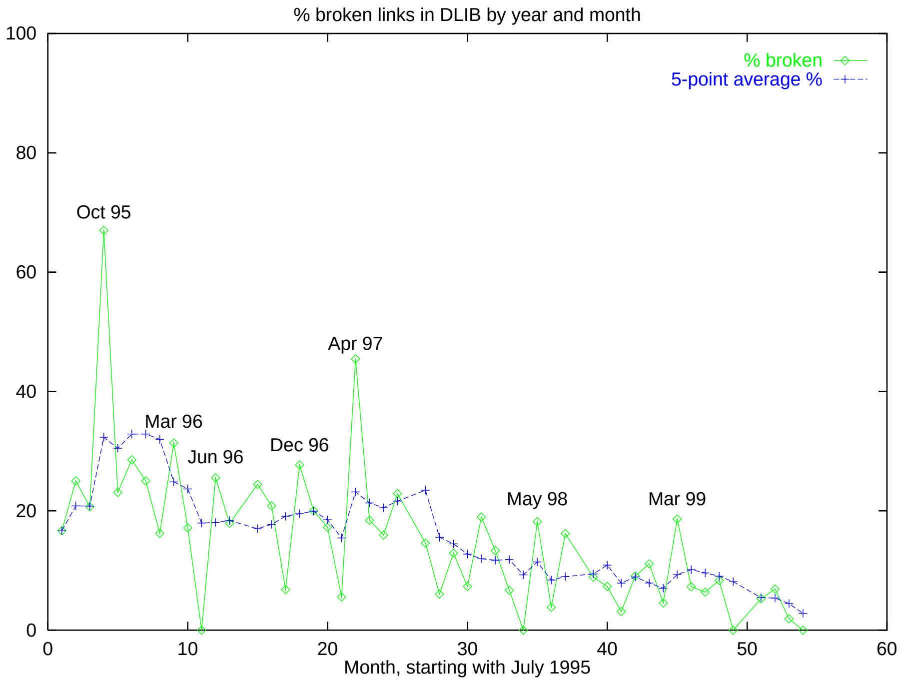

Figure 2: Links Broken and 10-point Average

Magazine. But as a sanity check we also looked at another well known online magazine, Journal of Electronic Publishing. The overall percentage of links broken over the five years of its life is very similar to D-Lib's: 13.36 %. The average number of links per article was 6.

This analysis involved the use of robot software to traverse the online literature. One problem was that the tools had to be slightly modified in order to parse mis-typed HTML comments. Such comments are handled gracefully by most web browsers, but the online parsing tools are meant to parse totally legal HTML. The tools also had trouble parsing some `robot.txt` files. Other than that, the package was easily put together and ran correctly almost from the start. Thus web crawling is a reasonable way to analyze reference links.

6 Conclusions

The number of broken links in almost 5 years of an online magazine is lower than expected. Not surprisingly, the later D-Lib issues have fewer broken links than the earlier ones. Another way to look at this same conclusion is Figure 3, which plots the total number of links encountered in all articles since 1995, versus the number that are now broken. It is clear from this graph that the number of broken links grows much more slowly than the total number of links from all articles, and that the rate of increase decreases with time.

It is quite interesting to compare our (admittedly limited) results with the analysis done by Harter and Kim [4]. At that point in time, namely late 1995, over half the cited online references contained a URL; now (2000) that is standard, unless the reference is to a journal that appears both online and on paper. In that case, you just go to the online journal if you have access and if you can resolve the reference to a URL.

In 1995, Harter and Kim attempted access to 83 online references; only 51.8% of them were accessible a year later. Of the ones that included an URL, 66% were accessible. Our results show that over 5 years, more than 86% remain accessible, on the corresponding kind of references. Also there are many more online references today. These are encouraging results for those who wish to implement reference linking for all online literature, including those scholarly journals which lie somewhere between gray literature and published journal literature.

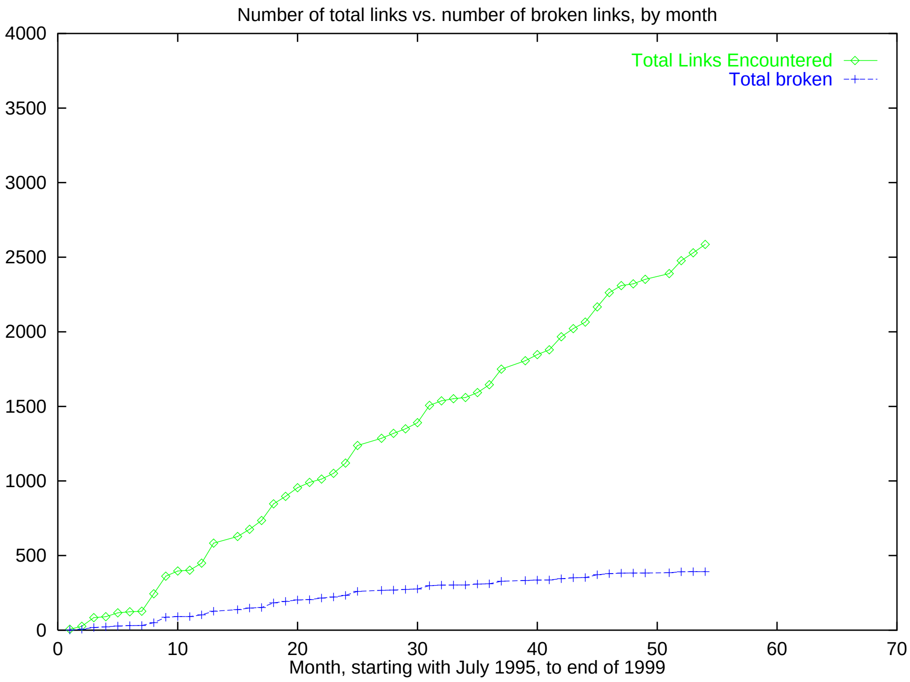

Figure 3: Cumulative Results

References

[1] CPAN, the comprehensive perl archive network. <http://www.perl.com/CPAN/>.

[2] William Arms. Preservation of scientific serials: Three current examples. *The Journal of Electronic Publishing*, May 1999.

[3] Lloyd A. Davidson and Kimberly Douglas. Promise and problems for scholarly publishing. *The Journal of Electronic Publishing*, 4(2), December 1998.

[4] Stephen P. Harter and Hak Joon Kim. Electronic journals and scholarly communication: A citation and reference study. In *Midyear Meeting of the American Society for Information Science*, pages 299–315, San Diego, CA, May 20-22 1996. <http://www.press.umich.edu/jep/archive/harter.html>.

[5] Steve Hitchcock, Les Carr, Wendy Hall, Stephen Harris, S. Probets, D. Evans, and D. Brailsford. Linking electronic journals: Lessons from the Open Journal project. *D-Lib Magazine: The Magazine of Digital Library Research* <http://www.dlib.org/dlib>, December 1998.

[6] Steve Hitchcock, Leslie Carr, and Wendy Hall. A survey of STM online journals 1990-05: The calm before the storm. Technical report, Electronic Libraries Program, Dept. of Electronics and Computer Science, University of Southampton, January 1996. updated February 1996. Available at <http://journals.ecs.soton.ac.uk/survey/survey.html>.

[7] Karen Hunter. Adding value by adding links. *The Journal of Electronic Publishing*, 3(1), September 1997. <http://www.press.umich.edu/jep/03-01/JEA.html>.

[8] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999. <http://www.researchindex.com>.

[9] Birdie MacLennan. Presentation and access issues for electronic journals in a medium-sized academic institution. *The*

Journal of Electronic Publishing, 5(1), September 1999. <`http://www.press.umich.edu/jep/05-01/maclennan.html`>.

[10] Norman Paskin. E-citations: actionable identifiers and scholarly referencing. 1999. <`http://www.doi.org/citations.pdf`>.

[11] Herbert Van de Sompel and Patrick Hochstenbach. Reference linking in a hybrid libary environment, part 1: Frameworks for linking. *D-Lib Magazine: The Magazine of Digital Library Research* <`http://www.dlib.org/dlib`>, 5(4), April 1999.

[12] Herbert Van de Sompel and Patrick Hochstenbach. Reference linking in a hybrid libary environment, part 2: Sfx, a generic linking solution. *D-Lib Magazine: The Magazine of Digital Library Research* <`http://www.dlib.org/dlib`>, 5(4), April 1999.

[13] Herbert Van de Sompel and Patrick Hochstenbach. Reference linking in a hybrid libary environment, part 3: Generalizing the sfx solution in the sfx@ghent & sfx@lanl experiment. *D-Lib Magazine: The Magazine of Digital Library Research* <`http://www.dlib.org/dlib`>, 5(10), October 1999.

[14] Herbert Van de Sompel and Carl Lagoze. The Santa Fe Convention of the Open Archives initiative. *D-Lib Magazine: The Magazine of Digital Library Research* <`http://www.dlib.org/dlib`>, 6(2), February 2000.