Abstract

The AusWeb conference has played a significant role in promoting and advancing research in web technologies in Australia over the last decade. Papers contributed to AusWeb are highly connected to the web in general, to digital libraries and to other conference sites in particular, owing to the publication of proceedings in HTML and full support for hyperlink based referencing. Authors have exploited this medium progressively more effectively, with the vast majority of references for AusWeb papers now being URIs as opposed to more traditional citation forms. The objective of this paper is to examine the reliability of URI citations in 10 years worth of AusWeb proceedings, particularly to determine the durability of such references, and to classify their causes of unavailability.

Introduction

The AusWeb conference has played a significant role in promoting and advancing research in web technologies in Australia over the last decade. In addition, the AusWeb forum serves as a point of reflection for practitioners engaged in web infrastructure, content and policy development; allowing the distillation of best practice in the management of web services particularly in higher educational institutions in the Australasian region. Papers contributed to AusWeb are highly connected to the web in general, to digital libraries and to other conference sites in particular, owing to the publication of proceedings in HTML and full support for hyperlink based referencing. Authors have exploited this medium progressively more effectively, with the vast majority of references for AusWeb papers now being URIs as opposed to more traditional citation forms. Although care has been taken by the proceedings editors to ensure that a high degree of syntactic standards compliance is adherred to by AusWeb authors, the editors are not responsible for paper content, or the citations made within them. As such, it is the responsibility of the authors (and arguably perhaps, the AusWeb reviewers) to ensure that the citations made in a given paper, particularly those cited by URI, are available for consideration by interested readers.

The objective of this paper is to examine the reliability of URI citations in 10 years worth of AusWeb proceedings, particularly to determine the durability of such references, and to classify their causes of unavailability. To our knowledge, the work in this paper is the only work specifically focusing on URI citation durability in conferences with web-based proceedings.

We consider the availability and persistence of URIs cited in the proceedings of AusWeb 1995-2005. 5975 unique URIs are extracted from 611 papers (referreed papers and edited posters) published in the 10 years worth of AusWeb proceedings available on the web at http://www.ausweb.scu.edu.au/. The availability of URIs was checked once per week for 6 months from September 2005 to March 2006. We found that approximately 42% of those URIs failed to resolve initially, and an almost identical number (42%) failed to resolve at the last check. A majority (99%) of the unresolved URIs were due to 404 (page not found) errors. We explore possible factors which may cause a URI to fail based on its age, path depth, top level domain and file extension. Based on the data collected we conclude that the half-life of a URI referenced in AusWeb proceedings papers is approximtely 6 years, and we compare this to previous studies of similar phenomena. We also find that URIs are more likely to be unavailable if they pointed to resources in the .net or .edu domains or country-specific top level domains, used non standard ports, or referenced a resource with an uncommon or deprecated file type extension.

Related Work

This study is derivative of a number of previous efforts in the study of the persistence of cited URIs in academic contexts. A summary can be found in the table below, prior to discussion in more detail.

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>URIs</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harter and Kim</td>
<td>1996</td>
<td>47</td>
<td>general scholarly journals</td>
</tr>
<tr>
<td>Koehler</td>
<td>1999</td>
<td>350</td>
<td>random URIs</td>
</tr>
<tr>
<td>Lawrence et al</td>
<td>2001</td>
<td>67,000</td>
<td>computer science pre-prints</td>
</tr>
<tr>
<td>Markwell and Brooks</td>
<td>2002</td>
<td>500</td>
<td>educational journals</td>
</tr>
<tr>
<td>Rumsey</td>
<td>2002</td>
<td>3,406</td>
<td>law journals</td>
</tr>
<tr>
<td>Nelson and Allen</td>
<td>2002</td>
<td>1,000</td>
<td>digital libraries</td>
</tr>
</tbody>
</table>
Table 1: Previous Similar Studies

Koelher (1999, 2004) provides arguably the longest continuous study of URI persistence using a set of around 350 URIs randomly selected in December 1996; finding that the "half-life" of a given URI is approximately 2 years. Another very early URI persistence study was undertaken by Harter and Kim (1996), who examined a small sample of 47 URIs from scholarly e-journals published between 1993 and 1995; their findings were that approximately 33% of the URIs were inaccessible within 2 years of their publication date. Other researchers (Markwell and Brooks, 2002) monitored around 500 URIs which referenced scientific or educational content in 2000-2001 and found that around 16% of their URIs became inaccessible or had their content change. Rumsey (2002) considered a far larger sample of 3406 URIs used in law review articles in 1997-2001 and found that 52% of URIs were not accessible at the time of the study in 2001. The persistence of 1000 digital objects using URIs from a collection of digital libraries was tested by Nelson and Allen (2002), although their context was considerably different in the sense of preservation awareness within a digital library and found a 3% loss of URIs during 2000-2001.

Two other studies have focused on persistence of URIs in computer science and related publications. Lawrence et al (2001) performed a study using around 67,000 URIs drawn from computer science papers in CiteSeer's Research Index that were published from 1993-1999. In May 2000, Lawrence et al accessed each of the 67,000 URIs and found that the percentage of invalid URIs increased from 23% in 1999 to 53% in 1994. Spinellis (2003) performed a similar experiment using around 4,200 URIs from computer science papers published in 1995-1999 by the ACM and IEEE Computer Society digital libraries. Measuring the URI availability in June 2000 and August 2002, Spinellis found that the half life of a referenced URI was approximately 4 years from the publication date, and also showed that URIs from older papers (1995-1996) aged at a significantly quicker rate than did URIs from later periods.

Most recently, McCown et al (2005) conducted a detailed analysis of the availability and persistence of URI references in D-Lib Magazine, using 4387 URIs. McCown et al found that the point at which more than half of the URIs are unavailable (the "half life") for D-Lib Magazine is around 10 years from date of publication, which they account for by virtue of D-Lib's editorial practice and the type of authors who write D-Lib papers (ie those with a bias towards URI persistence in some form).

To our knowledge, the work in this paper is the only work specifically focusing on URI citation durability in specifically web published proceedings of academic conferences. While we have a smaller sample size than that of eg Lawrence et al (2001), we have the advantage of a contiguous 10 years worth of conference papers from AusWeb. In both our methodology and relative sample size, our work is most directly comparable to that of McCown et al (2005); the key difference being that McCown et al (2005) focus on 10 years of an electronic journal, while we focus on 10 years of a conference proceedings.

Methodology

The core of our experiment is designed as follows:

1. From http://www.ausweb.scu.edu.au/aw06/archive/index.html we extracted all of the papers (referreed papers and edited posters) from AusWeb 1995 to AusWeb 2005, obtaining a total of 611 papers published over the course of approximately 10 years.
2. We downloaded all of the 611 papers.
3. From these 611 papers, we extracted all of the URIs; removing duplicate URIs per paper, and producing a total of 6178 URIs (an average of 10.1 per paper).
4. From this set of 6178 URIs we removed URIs which referenced AusWeb itself, and all redundant URIs producing a set of 5795 URIs.
5. We downloaded each of these 5795 URIs 26 times (once a week for 26 weeks) from September 2005 to March 2006, keeping track of the detailed response codes for each URI.

In determining redundant URIs, we considered that top level URIs which included and excluded a trailing / to be identical (eg http://www.ausweb.scu.edu.au and http://www.ausweb.scu.edu.au/). Below this level, the inclusion of a trailing / in a URI cannot be considered identical since it is possible for these URIs to resolve to different things.
We did not consider an index.html and a / to be equivalent, since although it is likely that they resolve to the same object, it is possible that they are different (eg by automatically generated directory content listings which are date and time stamped at the time of query). Similarly, we did not assume that http://www.ausweb.scu.edu.au/ and http://ausweb.scu.edu.au/ were equivalent despite widespread default web server configurations which resolve in this manner.

We additionally make no assumption about host and domain naming conventions. Although 77.6% of our URIs all begin with a 'www.', this is by no means universal.

We decided to exclude all the URIs which referenced the AusWeb server (http://www.ausweb.scu.edu.au) because the AusWeb pages are relatively actively maintained, a status which does not accurately reflect natural conditions on the web. Of the 6178 URIs we extracted, 198 (~3.2%) of them referenced http://www.ausweb.scu.edu.au pages and we found that around 98% of these were available on the first and last days of testing.

The retrieval module consists of a collection of Perl scripts based around the libwww CPAN library. An evaluation module consists of a series of Perl scripts which parse the retrieval module log files and collate various statistics. For each of the 26 retrieval events, we conducted the retrieval process 3 times and averaged the results across all retrieval runs for an event. The primary site from which these experiments were conducted is in Melbourne, Australia.

Results

Protocol

Figure 1 lists the distribution of retrieval protocols in the set of 5978 URIs. A vast majority of the URIs use HTTP, while all other protocols (including HTTPS) account for only a very small proportion of URIs. Notably, this list also includes several protocols which are deprecated (eg gopher) and others which are typically used incorrectly (eg file).

![Figure 1: URI Protocols](image)

For conducting our availability experiments, we excluded non HTTP URIs and syntactically malformed HTTP URIs from this point onwards, reducing the total number of URIs to 5795.

Overall

In total, 5795 URIs are cited in the AusWeb proceedings from 1995-2005, an average of 579 per year; or an average of 10.1 per paper.

Averaged over the retrieval intervals (26 retrievals), 56.05% of the 5795 URIs retrieved were still available. A total of 24.42% of the 5795 URIs consistently returned a 404 (not found) error. The remaining 19.53% of URIs returned fluctuating error types including 404 errors and others depending on the exact probe date and time.

Per Year URI Availability

Overall we can see an increasing availability of URIs with decreasing age; the lowest point is 27% of URIs from the 1995 AusWeb papers are available, with the high point being 85% of URIs from the 2005 AusWeb papers being available. A trend analysis shows an increase in availability of a cited URL of approximately 6% per year.

From our analysis, the half-life of a URI cited in the AusWeb proceedings is around 6 years; that is, 6 years after publication more than 50% of URIs are no longer available. The watershed year at the time of writing is 1999, at which 50% of URIs cited in the AusWeb proceedings are still available; that is, earlier than this date, less than half of the URIs cited will be retrievable.

![Figure 2: URIs and Availability Status by Year](image)

Error Code Typology

We recorded the HTTP status code for each of the URIs. By far the most common type of status code was 404, returned by 24.42% of all URIs, meaning that the resource no longer existed on the server. This error type made up 99% of the failures. Other failures we encountered in very small (and statistically insignificant) numbers were
400 (bad request), 401 (authentication) and 503 (service unavailable).

**Top Level Domain**

Figure 3 shows the distribution of URIs within our retrieval set based on their location within various parts of the global top level domain (gTLD) hierarchy. The majority (41%) are in the .edu name space, followed by 28% in the .com name space.

![Figure 3: Distribution of URIs by Global Top Level Domain (gTLD)](image)

In Figure 4 we can see the distribution of the URIs which fall within the .au ccTLD. Here, edu.au is the most common (65%), followed by other .au (19%) and then com.au (9%).

![Figure 4: Distribution of URIs within .au domain](image)

**Path Depth**

To determine how a URI’s path length influences decay rates, we followed the same method as McCown et al (2005) in adding 1 to the depth for every directory or file after the domain name, hence longer URIs have a greater numerical representation.

Table 2 shows the distribution of the 5795 URIs according to path depth. A significant number of the URIs have a path depth of 2 or less.

<table>
<thead>
<tr>
<th>Path Depth</th>
<th>% of URIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.4%</td>
</tr>
<tr>
<td>2</td>
<td>21.7</td>
</tr>
<tr>
<td>3</td>
<td>22.8%</td>
</tr>
<tr>
<td>4</td>
<td>26.4%</td>
</tr>
<tr>
<td>5</td>
<td>11.9%</td>
</tr>
<tr>
<td>6+</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

*Table 2: Path Depths of URIs*

We find that the average path depth of a URI is 3.1, that is, a URI will have around 3 directories after the host.domain portion. This value is greater than McCown et al (2005), who find that a majority of URIs in their set have a path depth of 2 or less.

**Path Characteristics**

Around 7.2% of URIs contain a ~, typically indicating a user's home directory on a Unix system. Of the 421 URIs which do contain a ~, around 68% of these URIs are no longer available; much higher than the average 56% across the entire set. This evidence corresponds with earlier findings that personal home pages referenced as URIs have a low level of reliability owing to the fluid relationship between individuals and entities, particular in educational contexts.

3.1% of our URIs contain a ?, taken by McCown et al (2005) to indicate a query string from which a dynamic page is produced. However we find that in contrast with McCown et al (2005) that these exhibit a greater reliability at least at the level of returning a non-error status. As discussed later, if we did utilise a content-checksum approach we could determine how often these dynamically generated URIs changed, if at all.

**MIME Types**

The vast majority of URIs referenced the text/htm or text/html MIME type (33.5% and 45.2% respectively). Of the remaining 21.3% binary document types such as pdf (3.3%), doc (9.9%) and ppt (4.7%) were apparent. Figure 5 provides more details of the MIME type distribution.
Our findings reinforce those of McCown et al (2005) in that the proportionally largest MIME type(s), htm and html, also incur the highest proportional unavailability. Interestingly, the pdf MIME type exhibited 100% availability within our set of URIs, indicating possibly that authors who are citing PDF documents are slightly more conscious about the durability of their URIs (alternatively, that the creators of the PDF documents and web server maintainers are more conscious of PDFs as a durable content type).

URI Relocation

We were able to capture data regarding the migration of URLs in various modes, by selecting particular types of HTTP information returned in response to queries. In particular, we find a significant number of URIs have migrated either temporarily (response type 301) or permanently (response type 302). 28.28% of the 5795 URIs returned either a 301 or 302 response type.

These figures are considerably advanced on previous studies for several reasons. of which most important is that we can consider both 301 and 302 errors as indicative of durability; that is, someone made a conscious decision to redirect HTTP requests for a given URI via either mechanism rather than simply letting the URI not resolve (hence a 404 error type).

TCP Service Ports

Only 160 of the 5975 URIs tested specified a TCP port (2.6%); the remainder are assumed to be TCP port 80 (the default HTTP port). Of those that did specify a TCP port, 50% of them specified the default (port 80); an additional 39 specified port 70 (40%). The remaining 10% specified an essentially random selection (1994, 591, 8000, 81, 801, 8001, 73). However, it is noticeable that the specification of non-standard ports is a feature of earlier years of the AusWeb proceedings, and in recent times (since 2001) there are zero instances of non-standard ports in cited URIs.

Discussion

We can compare our findings about cited URI half life with other previous studies. Our finding is that the half life of a URI cited in the AusWeb proceedings is 6 years. McCown et al (2005) found a longer half life within a particular publication (D-Lib Magazine), the half life of a referenced URI was 10 years. This longer half life can be accounted for by the fact that D-Lib is an actively preserved journal, and authors have a particular interest in digital preservation in various dimensions. Spinellis (2003) reported a shorter half life of 4 years, but drew sample data from a much earlier period (1995-1999). If we select the same period, the half life of AusWeb URIs also drops to around 4 years. Overall our figures are more akin to the McCown et al study since we gathered statistics over a much longer time frame (10 years, 1995-2005) than Spinellis (4 years, 1995-1999).

Our figures also differ from the McCown et al and Spinellis studies when considering path depth. We find that as path depth increases, so does the URI unavailability rate, which is similar to Spinellis' findings. However, we find that the rate at which unavailability increases is certainly greater than linear. McCown et al also found that path depth was an indicator of URI unavailability, but only for path depths up to 2, beyond which a levelling off effect was shown. We find that in our URI set at least, there is no levelling effect, but rather an increasing likelihood that URI failure would be encountered.

Our figures also show that AusWeb URIs have a longer half life than CiteSeer URIs reported by Lawrence et al (2001), but in similar fashion to McCown et al (2005) we believe that the early date of the data analysed in this work is not indicative of the last 10 years of web publishing and URI citation praxis. In keeping with McCown et al (2005), we extracted actual HREF links in the document source, rather than any occurrence of http, https or ftp in the text as followed by Lawrence et al (2001). We did find a small number of syntactically invalid URLs in our retrieval set; it should be noted that all of the file:// protocol URIs excluded earlier were in fact malformed.

We find a significantly lower decay rate than either Spinellis (2003) or Lawrence et al (2001). Our figures are around 5% decay in the first year, and then slightly higher for the next 3-4 years (6-7%). This contrasts with the Spinellis (2003) study which found a 20% initial decay, and then 10% per year thereafter; and the Lawrence et al (2001) study which found 23% initial decay and then 5% decay for 5 years. We believe the initial decay is significantly less than either of these earlier studies because of the nature of the domain: AusWeb ipso facto is concerned with web technologies, and as such authors arguably take greater care in URI citation.

Future Work
In this paper we have established a baseline methodology and measurements for the degree of URI citation durability in the AusWeb proceedings. Future editions of the AusWeb proceedings can be trivially added (helped by the fact that AusWeb publishes papers on the web in advance of the conference); and we intend to continue this study with the 2006 AusWeb proceedings.

One aspect of the McCown et al (2005) study which was not considered here was that of content persistence. To determine appropriate values, a secondary instrumentation framework is needed to measure the size of retrieved objects, and then a store of the object's previous size; neither of which is particularly onerous in a programmatic sense. However, based on manual inspection of a random 5% (290) of the URIs we retrieved, we believe that the proportion of URIs which vary with regard to size in the AusWeb proceedings case is likely to be very low: unlike earlier studies AusWeb authors have a tendency to refer to specific pages, rather than the generic references featured in earlier papers on the subject, and to refer to static, rather than dynamically generated URIs. Hence we did not believe that an additional investigation into these types of content changes was warrantee at this time, but could be an item for future work. Leveraging existing open source tools such as Bibcheck (Peacock, 1998) would natural way to achieve this end.

Another aspect of earlier work which was omitted here was that of the evaluation of persistence aware URIs: those which use handles, Permanent URLs (PURLS) or Digital Object Identifiers (DOIs). The reason for this omission to date is that none of the 5795 URIs found within the AusWeb proceedings actually referenced a resource by one of these means, and as such precluded any investigation. This is not to say that these types of URIs will not be cited in the future, and we should extend our analytical framework to account for their future occurrence.

Conclusion

This paper has considered the reliability of URI citations in 10 years worth of AusWeb proceedings, determining the durability of such references, and to classifying their causes of unavailability. As far as we are aware, no other work specifically focusing on URI citation durability in scholarly conferences with web-based proceedings; and we can correlate our findings against other studies in the general area with good effect.

We consider the availability and persistence of URIs cited in the proceedings of AusWeb 1995-2005. 5975 unique URIs are extracted from 611 papers (referreed papers and edited posters) published in the 10 years worth of AusWeb proceedings available on the web via http://www.ausweb.scu.edu.au/. The availability of URIs was checked once per week for 6 months from September 2005 to March 2006. We found that approximately 42% of those URIs failed to resolve initially, and an almost identical number (42%) failed to resolve at the last check. A majority (99%) of the unresolved URIs were due to 404 (page not found) errors. A number of possible factors which may cause a URI to fail based on its age, path depth, top level domain and file extension have been explored. Based on the data collected we conclude that the half-life of a URI referenced in AusWeb proceedings papers is approximately 6 years, which is placed around the median point of other relevant studies. We note that the decay rate for cited URIs is significantly less than in other comparable contexts. We also find that URIs are more likely to be unavailable if they pointed to resources in the net or edu domains or country-specific top level domains, used non standard ports, or referenced a resource with an uncommon or deprecated file type extension.

References


Link? Rot. URI Citation Durability in 10 Years of AusWeb Proceedings.


**Hypertext References**

HREF1

HREF2
http://www.csse.unimelb.edu.au/

HREF3
http://www.unimelb.edu.au/

HREF4
http://informationr.net/ir/2-1/paper9a.html

HREF5

HREF6
http://www.iwaw.net/05/papers/iwaw05-mccown1.pdf

HREF7
http://www.dlib.org/dlib/january02/nelson/01nelson.html

HREF8
ftp://ftp.ukoln.ac.uk/metadata/tools/bibcheck.cgi

**Copyright**

Baden Hughes, © 2006. The authors assign to Southern Cross University and other educational and non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to Southern Cross University to publish this document in full on the World Wide Web and on CD-ROM and in printed form with the conference papers and for the document to be published on mirrors on the World Wide Web.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s: 
HUGHES, BADEN

Title:
Link? Rot: URI citation durability in 10 years of AusWeb Proceedings

Date:
2006

Citation:

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/34215

File Description:
Link? Rot. URI citation durability in 10 Years of AusWeb Proceedings

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.