Analyzing the Persistence of Referenced Web Resources with Memento

Robert Sanderson
Mark Phillips
Herbert Van de Sompel

http://mementoweb.org/

Memento is funded by The Library of Congress
Overview

• Motivating Horror Story

• Memento

• Experiment

• Results

• Conclusions/Future Work
A Motivating Academic Horror Story

The long-term preservation of Web content

Michael Day
UKOLN, University of Bath
m.day@ukoln.ac.uk

1. Introduction

Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users. To date, most such initiatives have concentrated on the development of strategies and software tools for the collection of Web content and for providing current access to this content through interfaces like the Internet Archive’s Wayback Machine. The International Internet Preservation Consortium (IIPC) is currently building on this legacy with the collaborative development of a set of tools that can be used for the capture of Web sites and for the navigation and searching of Web archives. The focus on collection strategies and tools is a response to what is perhaps the most significant challenge of the Web from an information management perspective. Its dynamic nature means that pages, sites and even whole domains are continually evolving or disappearing.
A Motivating Academic Horror Story

The long-term preservation of Web content

Michael Day
UKOLN, University of Bath
m.day@ukoln.ac.uk

1. Introduction

Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users. To date, most initiatives have concentrated on the development of strategies and software tools for the collection of Web content and for providing current access to content through interfaces like the Internet Archive's Wayback Machine. The International Internet Preservation Consortium (IIPC) is currently building on this legacy with the collaborative development of a set of tools that will be used for the capture of Web archives. The focus is on what is perhaps the most difficult aspect of preservation: the management of Web sites and even whole domains.

1. Introduction

Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users. To date, most initiatives have concentrated on the development of strategies and software tools for the collection of Web content and for providing current access to content through interfaces like the Internet Archive's Wayback Machine. The International Internet Preservation Consortium (IIPC) is currently on this legacy with the collaborative development of a set of tools that can be used for the capture of snapshots of Web archives. The focus is now on what is perhaps the most important aspect of preservation, that of long-term access to and management of Web content and even whole domains.

A Motivating Academic Horror Story

The long-term preservation of Web content

Michael Day
UKOLN, University of Bath
m.day@ukoln.ac.uk

1. Introduction

Web archiving initiatives exist by current and future generations have concentrated on the development of the collection of Web content and through interfaces like the International Internet Preservation Consortium (IIPC) is currently building on this legacy with the collaborative development of a set of tools that can be used for the capture of Web archives. The focus now is perhaps the most information management platforms and even whole domains.

Another Motivating Academic Horror Story

July 2007

Link Rot: How the Inaccessibility of Electronic Citations Affects the Quality of New Zealand Scholarly Literature

Ailsa Parker

Whitireia Community Polytechnic, ailsa.parker@whitireia.ac.nz

Recommended Citation


http://www.coda.ac.nz/whitireia_library_jo/1
Another Motivating Academic Horror Story

Forbidden

You don't have permission to access /whitireia_library_jo/1/ on this server.

Apache/2.2.9 (Debian) Server at www.coda.ac.nz Port 80
Question 1

To what extent are web resources that are referenced from works in repositories still available at their original URL?

<table>
<thead>
<tr>
<th>Paper</th>
<th>URLs</th>
<th>Years</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casserly 2003</td>
<td>500</td>
<td>1999-2000</td>
<td>56.4% in 2002</td>
</tr>
<tr>
<td>Russell 2008</td>
<td>510</td>
<td>1999-2006</td>
<td>82% in 2006</td>
</tr>
<tr>
<td>Davis 2002</td>
<td>688</td>
<td>1999-2001</td>
<td>69% in 2002</td>
</tr>
<tr>
<td>Sellitto 2005</td>
<td>1043</td>
<td>1995-2003</td>
<td>54% in 2005</td>
</tr>
<tr>
<td>Dimitrova 2007</td>
<td>1126</td>
<td>2000-2003</td>
<td>61.3% in 2007</td>
</tr>
<tr>
<td>Parker 2007</td>
<td>1229</td>
<td>2002-2005</td>
<td>70% in 2007</td>
</tr>
<tr>
<td>Falagas 2007</td>
<td>1417</td>
<td>2003-2006</td>
<td>83% in 2006</td>
</tr>
<tr>
<td>Moghaddam 2010</td>
<td>1761</td>
<td>1995-2008</td>
<td>73% in 2010</td>
</tr>
<tr>
<td>Goh 2005</td>
<td>2516</td>
<td>1997-2003</td>
<td>69% in 2005</td>
</tr>
<tr>
<td>McCown 2005</td>
<td>4387</td>
<td>1995-2004</td>
<td>70% in 2005</td>
</tr>
<tr>
<td>Lawrence 2001</td>
<td>67577</td>
<td>1993-1999</td>
<td>75% in 2000</td>
</tr>
</tbody>
</table>

Significant prior art!

But very small scale other than Lawrence's early work on Citeseer

(See paper for references)
Our Hero Enters the Scene!
Question 1(redux)

To what extent are web resources that are referenced from works in repositories still available at their original URL … or from archives of web resources?

Prior art sketchy at best, as lacks automated method to enable discovery of archived web resources.
Memento Framework

Memento wants to make it easy to navigate the web of the past

- Global version indicator: Time
- Based on the primitives of the Web: resource, representation, content negotiation, link
- Functionality: Given a URI and a Datetime, resolve the closest archived copy
Original Resources and Mementos
Memento: Bridge from Present to Past
Multiple Archives

HTTP Link
timegate

Content Negotiation
Accept-Datetime = Tm

Content Negotiation
Accept-Datetime = Tn

Content Negotiation
Accept-Datetime = Ti

Content Negotiation
Accept-Datetime = Tj

URI-R
current
representation

Original
Resource

Memento
TimeGate

URI-M1

URI-M2

URI-M3

URI-M4

 Archived & Version Resources

Persistence of Referenced Web Resources
Open Repositories 2011, Austin TX, June 6-11
Original Resource’s Server Gone

Diagram showing the interaction between the original resource, a time gate, and archived resources with HTTP GET requests and accept datetime parameters.
Question 2

How long is the period between the publication of a paper and the archiving of a resource cited by that paper?

Memento allows us to answer this question.
Experiment

Using Memento, check all of the links extracted from papers in repositories to discover:

• Are they still resolvable at their Original URI?
• Are Mementos available in archives?
• What is the Memento-Datetime of the closest copy?

Data Set:

• University of North Texas Institutional Repository
  • 3595 works, 17965 unique URLs
  • May 1999 to August 2010
• arXiv
  • 400144 works, 144087 unique URLs
  • December 1993 to December 2009
• Total:
  • 162052 URLs, generating 306452 (URL, Paper) tuples
Experimental Process

- Extract Links
- Filter Links
- Normalize Links

- Extract Metadata
- Normalize Metadata

Results:

* We filter broken and intra/inter-repository links.

(URL, Paper, Time, Subject)
Results: Archiving Extent per Repository

UNT
- 72% in archives and/or still exist
- High proportion of archived URLs, possibly due to academic level and general disciplines

arXiv
- 78% in archives and/or still exist
- 45% still exist, but not archived! Possibly due to high value, but very discipline specific references
Results: Days between Publication and Archive

Typical long tail, but inexplicably similar curves at different scales for repositories.

arXiv: 45% within a month, 80% within a year
UNT: 48% within a month, 80% within a year
Results: Archiving Extent Per Discipline

**UNT**
- Most disciplines exhibit similar behavior, except History, Journalism and English with lower percentage archived.

**arXiv**
- Most disciplines exhibit similar behavior with very low percentage archived within one month, and very high percentage still dereferencable.
Conclusions

Difficulty of the Research:
• Getting set of URLs to check!
• Need access to full text … or …

Proposal for the Repository Community:

• Repositories should expose the links extracted from the full text of their resources
  • In metadata for the resource
  • In an Atom feed …

• To act as seed URL list for a (Memento compliant) web archive

• This would preserve the context of scholarly communication for future generations, in the same way repositories preserve the communication itself
Future Work

• Repeat with much larger dataset
  • JSTOR
  • CiteSeer
  • Astrophysics Data System
  • RePeC
  • PubMed
  • arXiv
  • 10+ ETD Repositories
  • SSRN (discussion ongoing)
  • Your repository?

• Investigate 45/80 similarity

• Community support for automated scholarly web archive project
Thank You!

- Rob Sanderson
  - Twitter: @azaroth42
  - Email: azaroth42@gmail.com
    or rsanderson@lanl.gov


- Slides: http://slidesha.re/

- Memento:
  - http://www.mementoweb.org/
  - http://groups.google.com/group/memento-dev