Link Rot + Content Drift

Peter Burnhill
EDINA, University of Edinburgh

for Hiberlink Team at University of Edinburgh & LANL Research Library

STM_2015
1 December 2015

Funded by the Andrew W. Mellon Foundation
Link Rot + Content Drift = Reference Rot

Peter Burnhill
EDINA, University of Edinburgh

for Hiberlink Team at University of Edinburgh & LANL Research Library

STM_2015
1 December 2015

Funded by the Andrew W. Mellon Foundation
‘Link Rot’ is known to be scary
Content Drift - is when content at end of URI has changed

(a) Over time, same URI pointed different (often unrelated) web pages

(b) Dynamic content as values on webpage changes over time

Not the same as when first seen and referenced by author
Reference Rot

“when links to web resources no longer point to what they once did”

This is **Threat** to **Integrity** of Scholarly Record.
This is report of ‘Hiberlink’ Investigation

1. Defined Reference Rot & the Threat it posed
2. Generated large-scale evidence to measure extent and way in which it exists & undermines the Scholarly Record
3. Then to envisage potential & practical Remedy

Project: Hiberlink

Funder: Andrew W. Mellon Foundation

Partners: University of Edinburgh
EDINA & Language Technology Group,
School of Informatics
Los Alamos National Laboratory
Research Library

2 years: March 2013 to June 2015
2. Generating Evidence via Large-scale Enquiry

c. 400,000 articles across three large corpora: ArXiv, PMC & Elsevier

This involved

• converting PDFs into XML
• locating references in the body of the text as well as under ‘References’
• extracting each and every URL
• then using a ‘white list’ of publisher websites:
  • Scholarly Web
  • Web-at-large

=> over a million web at large references
Focus on References to the Wild Web in Scholarly Communication

<table>
<thead>
<tr>
<th></th>
<th>=&gt; Scholarly Resources</th>
<th>=&gt; To Web at large</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link Rot</strong></td>
<td>DOI, HTTP version of DOI</td>
<td>‘Web today, gone tomorrow’</td>
</tr>
<tr>
<td><strong>Content Decay</strong></td>
<td>Has ‘fixity’</td>
<td>Need to add fixity to the dynamic</td>
</tr>
<tr>
<td></td>
<td>Archiving: CLOCKSS, Portico, LOCKSS, etc, as per Keepers Registry …</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://thekkeepers.org">http://thekkeepers.org</a></td>
<td></td>
</tr>
</tbody>
</table>

Focus for Hiberlink
Scholarly Articles increasingly link to Web Resources, not just back to other Articles.
We then began to ask questions of those URIs:

1. Is the URI still on the ‘Live Web’?
   - Allowed up to a maximum of 50 redirects

2. Is there a ‘Memento’ of that content in the ‘Archived Web’?
   - Internet Archive, archive.is (archive.today), Archive-It, BL Web Archive, UK National Archives Web Archive & Icelandic National Archive

Memento: a prior version, what the Original Resource was like at some time in the past.

3. Does what exists at the end of that URI correspond to the content that the author intended?
Results: Most Referenced URIs at risk, Many are Lost, within 14 days of publication date!


http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0115253
This is also true for what is commercially published: within 14 days of publication date!

3 in 4 are at risk of loss

A third are lost forever!

http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0115253
Andy Jackson (2015) Ten years of the UK web archive: what have we saved?
http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_03_Jackson.pptx
=> Content of Citations Rot over Time!!
... meaning rotten references for the reader
Rot in References means a Defective Article!

... and sale of rotten goods 😞
Remedy for fish: Quick Freeze & Archive
Remedy for References: Snapshots of Web Content

When best to intervene within 3 workflows:

① **Preparation** -> Study -> Compose -> Submission

② **Publication** -> Editing -> (Revision) -> Acceptance -> Issue

③ **Access platform/post-publication** -> Reader Access -> Use
Clearly best at the earliest moment of capture
… when the Authors are trawling for content
... for what an Author regards as significant
or needs to provide as evidence
3. Potential & Practical Remedy: Take a Snapshot & put in safe place until needed

a) Use web-scale archives that support on-demand creation of snapshots of URIs:
   – archive.today; Internet Archive; perma.cc; webcitation.org

b) Embed this action in software used into basic workflows:

<table>
<thead>
<tr>
<th>Activity</th>
<th>Actor</th>
<th>Snapshot Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preparation</td>
<td>Author/reference tool</td>
<td>best</td>
</tr>
<tr>
<td>2. Submission /Issue</td>
<td>Editor/manuscript system</td>
<td>good</td>
</tr>
<tr>
<td>3. Access (post-publication)</td>
<td>Aggregator/publisher platform</td>
<td>better late than not</td>
</tr>
<tr>
<td>4. Shelving</td>
<td>Librarian/IR, journal archive</td>
<td>better than nothing</td>
</tr>
</tbody>
</table>
‘Best’: help authors do the right thing!

1. Preparation -> Study -> Compose -> Submission

a) We put focus on note-taking software: eg EndNote, Mendeley, Reference Manager, RefMe, Zorero

b) We developed Plug-in for Zotero [open source]

  • So good things could happen under the hood!
Actionable Metadata: **Use a ‘Robust Link’**

- Robust links are modified `<a>` HTML elements
  
  a) Take simple URI - to article in New Yorker magazine (say)
  
    - `http://www.newyorker.com/magazine/2015/01/26/cobweb`

  b) Augment Link with Archive URI and Datetime

    - Archive timestamp: 2015-02-19T09:46:36

  => & so construct & cite the Robust Link:


---

Herbert Van de Sompel et al. (2015) Robust Links - Link Decorations
http://robustlinks.mementoweb.org/spec/
What should we expect of the Publisher?

Beyond the assurance that the fish / references / articles sold are not rotten
How Can Publishers Stop/Halt Reference Rot?

• You are better placed to know that!
• But here are some Hiberlink suggestions …

At the Point of Ingest
How Can Publishers Stop/Halt Reference Rot?

• You are better placed to know that!
• But here are some Hiberlink suggestions …

At the Point of Sale & Use
## Recommended Actions on Reference Rot

<table>
<thead>
<tr>
<th>Activity</th>
<th>Responsibility</th>
<th>Snapshot Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Submission/Issue</td>
<td>Editor/System</td>
<td>good</td>
</tr>
<tr>
<td>3. Access</td>
<td>Platform</td>
<td>better late than not</td>
</tr>
</tbody>
</table>

### 2. Publication -> Editing -> (Revision) -> Acceptance -> Issue
- Accept Robust Links in Cited References!
- Batch archive snapshots & use Robust Links

### 3. Access/Post-Publication -> Reader Access -> Use
- Employ ‘Link Decoration’ & Robust Links for references in past publications
## Link Decoration: JavaScript + Memento API

### Reference List

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Resolve a DOI Name <a href="http://dx.doi.org">http://dx.doi.org</a> Accessed: 1 Nov 2014</td>
</tr>
</tbody>
</table>

---

**Demo** - [http://robustlinks.mementoweb.org/demo/uri_references_js.html](http://robustlinks.mementoweb.org/demo/uri_references_js.html)

**robustlinks.js** - [https://github.com/mementoweb/robustlinks](https://github.com/mementoweb/robustlinks)
Creating Robust Links ‘post hoc’ …

If no ‘date accessed’, then use article’s Publication Date, or *better still*, the Acceptance Date

- Express the Date in an actionable manner (‘datePublished’ or ‘dateModified’ Schema.org properties) in HTML pages that contain URI references

- Tailor robustlinks.js to exclude links to articles

- Inject robustlinks.js in HTML pages that contain URI references

… enables Users to follow that Link into Web Archives using Memento and/or WayBack Machine
Be Proactive: Referenced Web Content for new articles

When ingesting new content into the platform:

– Parse for URI references
– Separate references to web-at-large from publisher sites
– Create snapshots in web archives of those URIs
– Then use Link Decorations in HTML to convey:
  • original URI + snapshot URI + snapshot Date/Time
Our remedy was to write plugin for OJS

1. A parser converts .pdf to .html & extracts URIs

2. Triggers archiving of content for each reference
   - Author & Editor need to work together to determine which archival copy is used

3. Creates an HTML version that includes the Robust Link for each cited reference.

Algorithm for OJS plugin should generalise to other submission systems
We have also sketched HiberActive Infrastructure to act as middleware between existing software & web archives

- **Asynchronous** (returns Robust Link)
- **Distributed** (archived with different organisations)
- **Lightweight** (leveraging HTTP & what already exists)
Summary: Hiberlink Outcomes & Next Steps

1. Defined the **Threat** of Reference Rot
2. **Quantified** the extent and way in which it exists & undermines the Scholarly Record
3. Pointed to potential & practical **Remedy**
4. Tell the world about these achievements
5. Engage with others
   - to build infrastructure
   - to prompt adoption (copying) of prototypes by 3rd parties, such as reference managers, editorial systems, publication systems, archival systems
Published References have Robust Links to what the author intended 😊
Thank you,
Questions welcome
& any interest in working together

Email: edina@ed.ac.uk