Ensuring the Integrity (& Continuity) of Our Record of Scholarship

+ 7 Good Things To Do

Peter Burnhill
EDINA, University of Edinburgh

‘Standing on the Digits of Giants: Research data, preservation and innovation’

ALPSP/DPC, London 8th March 2016

Funding from the Andrew W. Mellon Foundation
How brittle are those Digits (of Giants) we presume to stand upon?

Focus on Scholarly Statement = Content + References to Content

<table>
<thead>
<tr>
<th>Content</th>
<th>Scholarly Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>References to Content</td>
<td>⇒ Back into Scholarly Publications</td>
</tr>
<tr>
<td></td>
<td>Has ‘fixity’</td>
</tr>
<tr>
<td></td>
<td>DOI, ISSN CLOCKSS, Portico, LOCKSS, etc</td>
</tr>
</tbody>
</table>

E-Journal Archiving

#keepers
How brittle are those Digits (of Giants) we presume to stand upon?

Focus on Scholarly Statement = Content + References to Content

<table>
<thead>
<tr>
<th>Content</th>
<th>Scholarly Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>References to Content</td>
<td>=&gt; Back into Scholarly Publications</td>
</tr>
<tr>
<td></td>
<td>Has ‘fixity’</td>
</tr>
<tr>
<td></td>
<td>DOI, ISSN CLOCKSS, Portico, LOCKSS, etc</td>
</tr>
</tbody>
</table>

E-Journal Archiving #keepers

Reference Rot #hiberlink
Scholarly Articles increasingly link to Wild Web Resources not just back to other Articles

- arXiv
- PMC
- Elsevier corpus

Dark solid lines represent URIs to Web-at-large, from 1997/2011

* Artefact of XML supplied

http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0115253
The Scholarly Record always had a fuzzy edge... including what is cited on the World Wide Web.
Unintended Consequences of The Web/Internet:
Digital back copy no longer in the custody of libraries

Picture credit: http://somanybooksblog.com/2009/03/27/library-tour/
Unintended Consequences of The Web/Internet:
Digital back copy no longer in the custody of libraries

Libraries boast of ‘e-collections’,
but do they only have ‘e-connections’?

Picture credit: http://somanybooksblog.com/2009/03/27/library-tour/
Good News: We do have some digital shelving 😊

- Web-scale not-for-profit archiving agencies:
  +
- National institutions (usually national libraries) …
  +
- Consortia of university libraries & specialist centres …

We now have means to discover who is looking after what e-seriais, via the Keepers Registry

"Tales from the Keepers Registry” Serials Review 39.1 (2013)
The Keepers Registry gives evidence on progress content from 30,769 titles being ingested with archival intent by > 1 keeper.
Key Statistics using Titles Ingested / Titles with ISSN

Ingest Ratio = titles ingested by one or more Keeper / total ‘online serials’ in ISSN Register
= \( \frac{30,769}{177,631} \)
=> 17%

KeepSafe Ratio = with 3+ Keepers / ISSN Register
= \( \frac{11,312}{177,631} \)
=> 6%

Total number of Online Serials in ISSN Register has been increasing: 177,631 in January 2016
### Big Variation in Archival Status of Online Continuing Resources (assigned ISSN) by Country, July 2015

<table>
<thead>
<tr>
<th>Country</th>
<th>ISSN Count</th>
<th>Ingest Ratio %</th>
<th>CLOCKSS or Portico %</th>
<th>A National Institution %</th>
<th>Other Agency %</th>
<th>KeepSafe Ratio %</th>
<th>Archival status unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>Elsevier 5411</td>
<td>68</td>
<td>60</td>
<td>54</td>
<td>54</td>
<td>49</td>
<td>1729</td>
</tr>
<tr>
<td>Egypt</td>
<td>Hindawi 1092</td>
<td>60</td>
<td>59</td>
<td>3</td>
<td>23</td>
<td>4</td>
<td>438</td>
</tr>
<tr>
<td>UK</td>
<td>T&amp;F, OUP, etc 14569</td>
<td>45</td>
<td>34</td>
<td>15</td>
<td>34</td>
<td>20</td>
<td>8066</td>
</tr>
<tr>
<td>USA</td>
<td>Wiley etc 31757</td>
<td>34</td>
<td>17</td>
<td>5</td>
<td>28</td>
<td>7</td>
<td>20911</td>
</tr>
<tr>
<td>Germany</td>
<td>Springer Karger 7121</td>
<td>27</td>
<td>18</td>
<td>11</td>
<td>18</td>
<td>10</td>
<td>5228</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1688</td>
<td>16</td>
<td>12</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>1421</td>
</tr>
<tr>
<td>Sweden</td>
<td>1015</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>01</td>
<td>911</td>
</tr>
<tr>
<td>Poland</td>
<td>2693</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2457</td>
</tr>
<tr>
<td>Canada</td>
<td>8655</td>
<td>8</td>
<td>2</td>
<td>0.2</td>
<td>6</td>
<td>0.2</td>
<td>7988</td>
</tr>
<tr>
<td>Russia</td>
<td>12118</td>
<td>7</td>
<td>6</td>
<td>1.2</td>
<td>3</td>
<td>3</td>
<td>4929</td>
</tr>
<tr>
<td>India</td>
<td>5248</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4899</td>
</tr>
<tr>
<td>Korea S</td>
<td>1071</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>1002</td>
</tr>
<tr>
<td>Japan</td>
<td>2187</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>2050</td>
</tr>
<tr>
<td>Italy</td>
<td>3576</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3368</td>
</tr>
<tr>
<td>France</td>
<td>12118</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>11538</td>
</tr>
<tr>
<td>Australia</td>
<td>5078</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>4846</td>
</tr>
</tbody>
</table>

---

Mostly Big Publisher content that’s being archived.
Standing on the Digits of Giants?
- Not If we don’t keep the digits ...

Priority:
find economic way to archive content from
very many ‘at risk’ e-journals from many (small & not so small) publishers
So, First Good Thing To Do – today/now

① Go to the Keepers Registry => thekeepers.org

✧ Search on Title/ISSN
  • Check key volumes & issues are being archived
✧ Browse by publisher
✧ Use the Title List Comparison tool [Member Services]
  • Are your Titles being archived?
✧ Consider the Linking Options to display ‘archival status’ for each Title on your website
There is Threat to **Integrity** of our Scholarly Record

“when links to web resources no longer point to what was intended”

**Reference Rot** = Link Rot + Content Drift

Funding: Andrew W. Mellon Foundation
Link Rot’ is known to be scary
Content Drift may be even scarier!
When what is at end of cited URL has changed, or gone!!

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

(b) Static content but very different (often unrelated) web pages
Hiberlink analysed 1 million URI links to Web-at-large not links to publisher & access platforms (DOI etc)

<table>
<thead>
<tr>
<th></th>
<th>URI references</th>
<th>arXiv</th>
<th>Elsevier</th>
<th>PMC</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>total</td>
<td>781,895</td>
<td>1,548,496</td>
<td>1,653,567</td>
<td>3,983,958</td>
</tr>
<tr>
<td>2</td>
<td>excluded</td>
<td>1,555</td>
<td>1,293,191</td>
<td>428,036</td>
<td>1,722,782</td>
</tr>
<tr>
<td>3</td>
<td>to articles</td>
<td>434,163</td>
<td>22,593</td>
<td>744,678</td>
<td>1,201,434</td>
</tr>
<tr>
<td>4</td>
<td>to web at large</td>
<td>346,177</td>
<td>232,712</td>
<td>480,853</td>
<td>1,059,742</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0115253.t003

**Methodology:** answer to 2 questions

1. Do those links (URIs) still work? - on the ‘Live Web’?
2. Is there a ‘Memento’ of that reference in the ‘Archived Web’?

If a Memento cannot be found in a Web Archive within N days of the date of publication, but URI is still active then risk of loss (& rot)

If Memento cannot be found in a Web Archive within N days of the date of publication, and URI not active on the Live Web, then it is lost / rotten
Methodology: answer to 2 questions

1. Do those links (URIs) still work? - on the ‘Live Web’?
2. Is there a ‘Memento’ of that reference in the ‘Archived Web’?
Hiberlink Results: *within 14 days of publication date* ...

<table>
<thead>
<tr>
<th></th>
<th>PMC</th>
<th>Elsevier</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Not Archived’</td>
<td>74.5%</td>
<td>75.2%</td>
</tr>
</tbody>
</table>

Most referenced URLs at risk of loss

### Of those ‘Not Archived’

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>still ‘Live’ on the Web</td>
<td>80</td>
<td>67.3</td>
</tr>
<tr>
<td>‘No longer Live’ on the Web</td>
<td>20%</td>
<td>32.7%</td>
</tr>
</tbody>
</table>

Reference Rot is already significant

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.
PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253

Team at Harvard Law School establishing similar evidence

“We documented a serious problem of reference rot:

- more than 70% of the URLs within the above mentioned [law] journals, and
- 50% of the URLs within U.S. Supreme Court opinions suffer reference rot — meaning, again, that they do not produce the information originally cited.”

Jonathan Zittrain, Kendra Albert and Lawrence Lessig (2014).
Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations.
=> Content of Citations Rot over Time!!

... leading to rotten references for the reader
Rot in References means a Defective Article!

undermines the integrity of the scholarly record 😞
Things To Do?

Hint: Remedy for fish is ‘Quick Freeze & Store’
What should the Publisher do?

To give assurance that the fish / references / articles sold are not rotten!
Help the Reader at the Point of Sale & Use with ‘Link Decoration’: JavaScript + Memento API

Reference List

2. Resolve a DOI Name http://dx.doi.org/ Accessed: 1 Nov 2014

Demo - http://robustlinks.mementoweb.org/demo/uri_references_js.html
robustlinks.js - https://github.com/mementoweb/robustlinks

… Reader is then taken to web content nearest in time to the submission date of the article
But best done at earliest moment of capture
... for what an Author regards as significant
... or needs to provide as evidence
More Things To Do: ‘Hiberlink Remedy’

③ **Snapshot & Save: Proactive/Transactional archiving**
- Use web-scale archives for on-demand snapshots of URIs:
  - archive.today; Internet Archive; perma.cc; webcitation.org

④ **Turn a simple URI - to article in New Yorker magazine (say)**
http://www.newyorker.com/magazine/2015/01/26/cobweb

into a hiberlink URI

```html
<a href="http://www.newyorker.com/magazine/2015/01/26/cobweb"
data-versiondate="2015-02-19T09:46:36">Cobweb</a>
```

Snapshot URI + Original URI + Date/Time  [Robust Link syntax]
http://robustlinks.mementoweb.org/spec/
Author workflow: note-taking software
eg EndNote, Mendeley, Reference Manager, Zotero, RefME

Semi-automatic archiving of referenced web content when noted

Software stores those ‘hiberlinks’ for use in citations
Two Things Publishers should do in Workflow?

Submission -> Editing -> (Revision) -> Acceptance -> Issue

⑤ Accept Robust Links in Cited References!

⑥ Avoid reference rot by triggering archiving of snapshots & inserting Hiberlinks / Robust Links at the point of Ingest in the submission system

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 the archival copy to use
3. HTML version that includes Robust Link for each cited reference.
Engage with us in HiberActive Infrastructure * to act as middleware between existing software & web archives

- Asynchronous, returns hiberlink in Robust Link syntax
- Distributed, enables archiving with different web archives
- Lightweight, leverages HTTP & what already exists

* In development
Standing on the Digits of Giants?
- only if published references have Robust Links to what the author intended

Questions welcome
& any interest in working together

edina@ed.ac.uk