Robust Links in Scholarly Communication

Martin Klein
Los Alamos National Laboratory
Los Alamos, NM, USA
http://orcid.org/0000-0003-0130-2097
mklein@lanl.gov

Harihar Shankar
Los Alamos National Laboratory
Los Alamos, NM, USA
http://orcid.org/0000-0003-4949-0728
harihar@lanl.gov

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, NM, USA
http://orcid.org/0000-0002-0715-6126
herbertv@lanl.gov

ABSTRACT

Web resources change over time and many ultimately disappear. While this has become an inconvenient reality in day-to-day use of the web, it is problematic when these resources are referenced in scholarship where it is expected that referenced materials can reliably be revisited. We introduce Robust Links, an approach aimed at maintaining the integrity of the scholarly record in a dynamic web environment. The approach consists of archiving web resources when referencing them and decorating links to convey information that supports accessing referenced resources both on the live web and in web archives.

KEYWORDS

scholarly communication, link rot, content drift, persistence, persistent identifiers, web archiving

ACM Reference Format:
Martin Klein, Harihar Shankar, and Herbert Van de Sompel. 2018. Robust Links in Scholarly Communication. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203885

1 INTRODUCTION

Our previous research [2] revealed that, increasingly, scholarly articles contain URI references to web resources such as project websites, online debates, presentations, blogs, videos, etc. Just like any other resource on the web, these referenced resources are subject to link rot and content drift, the combination of which is referred to as reference rot. As a result, over time, it becomes impossible to revisit these resources when reading a scholarly work in which they are referenced [1, 2]. These web resources that are referenced in scholarship are distinct from referenced journal articles in two ways:

(1) **Archiving:** They are not systematically archived, as is the case with journal articles through efforts such as LOCKSS[1], CLOCKSS[2], and Portico[3].
(2) **Linking:** The core assumption on which the use of persistent identifiers (PIDs) is based does not hold true for these web resources. Indeed, PID solutions rely on the custodians of PID-identified resources to keep links to their content operational. The incentive for these custodians to invest in maintaining link stability is significant because working links are crucial for accessing their collections. The creators of referenced web resources do not share that motivation. They are typically not overly concerned about the longevity of their own web resource, and definitely not about the integrity of the scholarly record. As such, a PID approach to link stability is not a realistic proposition as the custodian of the linked resource can not be expected to keep PID-based links operational.

In this poster, we present Robust Links[4], an approach aimed at increasing the integrity of the scholarly record, when references to web resources are concerned. The Robust Links concept consists of two components:

(1) **Archiving:** When referencing a web resource, pro-actively archive it using existing web archiving infrastructure.
(2) **Linking:** When linking to a web resources, decorate the link to convey information that supports to seamlessly visit the archived version of the resource when the resource itself has disappeared from the live web or when its content has drifted.

We detail both components in the following sections.

2 PRO-ACTIVE ARCHIVING

In order to make sure that referenced resources can reliably be revisited, a Memento (snapshot) of the resource should be created in one or more web archives. Ideally, this step is taken during the authoring process, while the author inspects a resource and decides to reference it. This could, for example, happen automatically when bookmarking a web resource in a reference manager[5]. But other scenarios are conceivable, such as archiving referenced resources in bulk upon submission of a manuscript to a journal/conference submission system or upon publication of the article[6]. However, our research regarding reference rot has shown that the more time has passed since authoring and referencing, the more likely it is that the referenced resource is suffering from reference rot. Therefore, the sooner a Memento of a resource is created after it was referenced, the more representative that Memento will be of what the author actually intended to reference.

[1] https://www.lockss.org/
[2] https://clockss.org/
[3] https://www.portico.org/

[4] http://robustlinks.mementoweb.org/
[5] http://hiberlink.org/zotero.html
[6] https://www.slideshare.net/martinklein0815/hiberactive

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203885

Special-purpose archives such as perma.cc[7], WebCite[8], and weblock.io[9] are stepping into this problem domain and are exploring avenues to meet the challenge of operating web archives for the long term. Other established web archives such as the Internet Archive[10] and archive.is[11] also provide solutions to pro-actively create Mementos of web resources.

3 LINK DECORATION

Once a Memento is created, it would seem logical to reference the Memento rather than the live web resource. However, doing so assumes that the archive in which the Memento resides will exist forever. When the archive ceases to exist, the link to the Memento will be rotten. Also, when merely linking to the Memento, it becomes impossible to visit the live web resource. Link decoration aims at providing as many pathways as possible to visit versions of a referenced resource. The link decoration is done using HTML5 extension attributes that start with `data-`. The regular link, with the original URI of the reference in `href` is augmented with the URI of the Memento (`data-versionurl`) and the datetime of linking (`data-versiondate`). The example below shows a Robust Link to the JCDL 2018 conference website.

```
<a href="https://2018.jcdl.org/"
   data-versionurl="http://archive.is/WLgTv"
   data-versiondate="2018-02-01">JCDL 2018</a>
```

These three information elements suffice to make a reference robust:

- The original URI supports revisiting the referenced resource as it evolves over time.
- The Memento URI supports accessing the Memento that was pro-actively created.
- The combination of the original URI and the datetime of linking supports accessing a Memento of the referenced resource with an archival datetime close to the linking datetime, in any web archive, either manually or using the Memento infrastructure[12].

Use cases exist where it is appropriate to use the Memento URI as the link target (`href`) and the original URI as a link decoration (`data-originalurl`). For example, a publisher may prefer to consistently link to Mementos rather than live versions of web resources that might, over time, return the infamous "404 - Page not found". The example below shows such a Robust Link, again, for the JCDL website.

```
<a href="http://archive.is/WLgTv"
   data-originalurl="https://2018.jcdl.org/"
   data-versiondate="2018-02-01">JCDL 2018</a>
```

4 ACTIONABLE ROBUST LINKS

By utilizing legitimate HTML5 extension attributes, the above link decoration approach ensures that the information is conveyed in an interoperable machine-actionable manner. Applications (client and server) can process decorated links to offer additional functionality

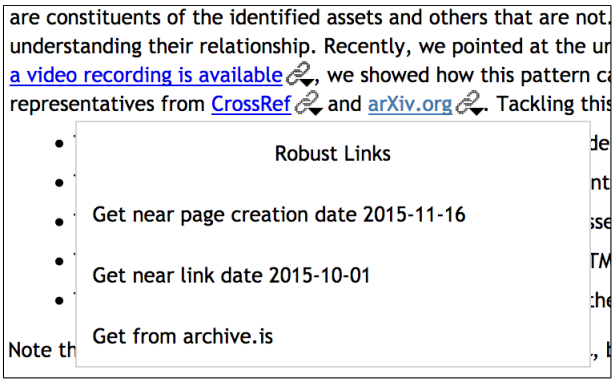

Figure 1: Actionable Robust Links

to the user. Figure 1 shows a screenshot of a section of a scholarly article [3] published in D-Lib Magazine that contains several Robust Links. These were made actionable by a few lines of JavaScript[13] embedded in the article. The JavaScript introduces the following navigation options:

- to access a Memento created in any archive closest to the date the article was published (top option),
- to access a Memento created in any archive closest to the date the link was created (middle option), and
- to access its Memento specifically created with the pro-active archiving service `archive.is` (bottom option).

The first option results from processing the page publication date conveyed in the paper's HTML:

```
<meta itemprop="datePublished" content="2015-11-16" />
```

while the second and third options result from the link decorations.

5 CONCLUSION

Links on the web are brittle. Archiving linked resources and decorating links are ways to make links more robust. Applications can leverage the link decorations to provide navigational paths to various versions of a referenced resource. Preferably, this includes linking to a Memento that, ideally, was created temporally close to the time of referencing the resource. The functionality that can be provided by adopting the Robust Links approach is attractive for the web in general. However, it is especially relevant for web-based scholarly communication because it helps to increase the integrity of the scholarly record.

REFERENCES

[1] Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover. 2016. Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PLoS ONE* 11, 12 (2016). https://doi.org/10.1371/journal.pone.0167475

[2] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE* 9, 12 (2014). https://doi.org/10.1371/journal.pone.0115253

[3] Herbert Van de Sompel and Michael L. Nelson. 2015. Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine* 21, 11/12 (2015). https://doi.org/10.1045/november2015-vandesompel

[7] https://perma.cc/
[8] https://www.webcitation.org/
[9] https://weblock.io/
[10] http://web.archive.org/
[11] http://archive.is/
[12] https://tools.ietf.org/html/rfc7089
[13] https://github.com/mementoweb/robustlinks