


Reference rot in web-based scholarly communication and link decoration as a path to mitigation

 blogs.lse.ac.uk/impactofsocialsciences/2015/02/05/reference-rot-in-web-based-scholarly-communication/

2/5/2015

The failure of a web address to link to the appropriate online source is a significant problem facing scholarly material. [Martin Klein](#) and [Herbert Van de Sompel](#) together with their collaborators have investigated the extent of reference rot on scholarly domains and their results show an alarming link rot ratio. The authors also explore ways to mitigate it through more systematic web archiving practices and link decoration techniques.



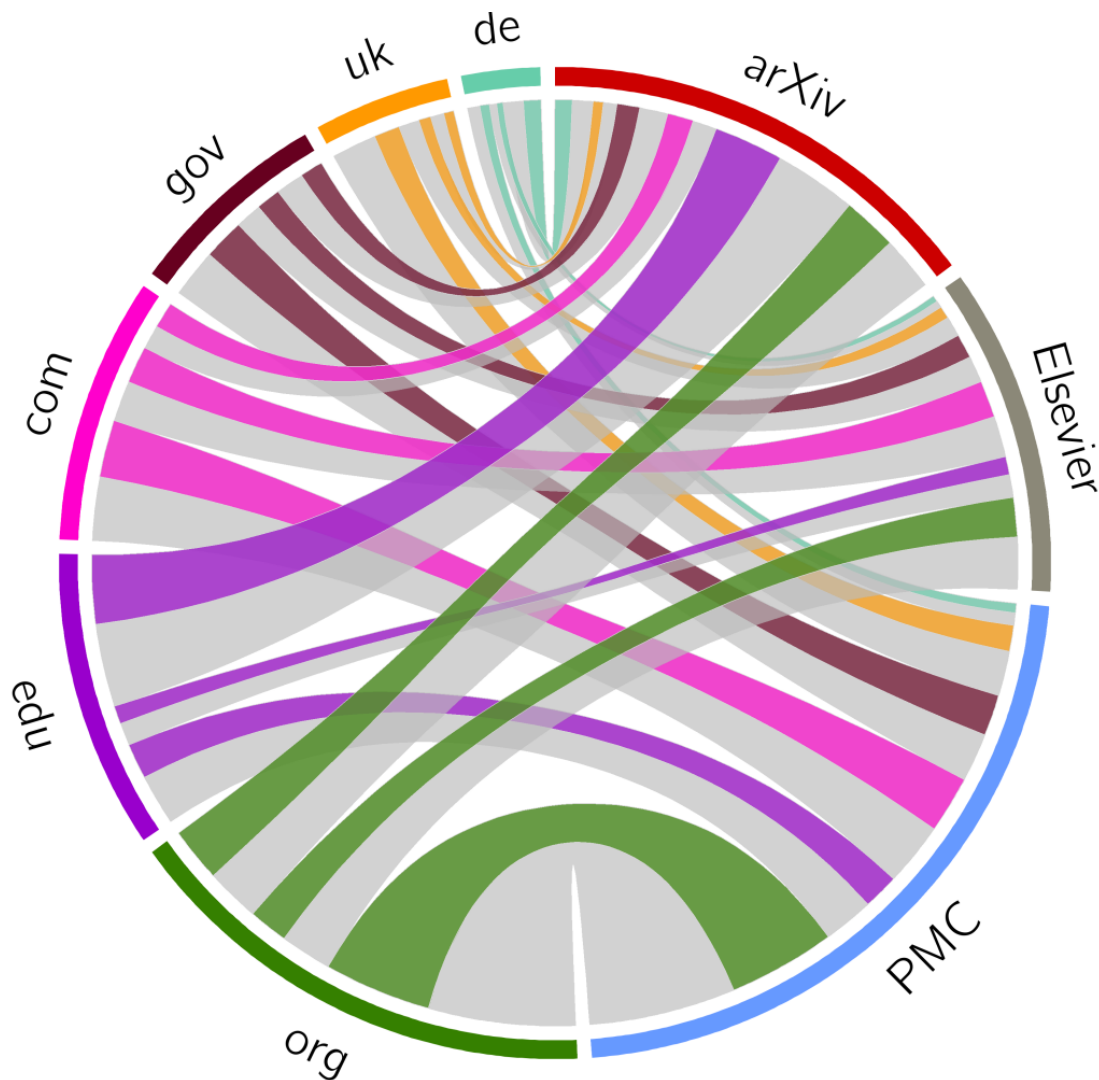
Referencing sources is a fundamental part of the scholarly discourse. This was true 50 years ago and still holds true today, even though, with the emergence of the web, scholarly communication has undergone a dramatic shift from a paper-based to a web-based endeavor. What has changed, however, is the variety of scholarly assets being referenced. Increasingly, we see references to software, ontologies, project websites, presentations, blogs, videos, tweets, etc. Such resources are usually referenced by means of their HTTP URI as they exist on the web at large. These HTTP URIs allow for immediate access on the web, but also introduce one of the detrimental characteristics of the web to scholarly communication: reference rot.



We introduced the term reference rot as a combination of two problems common for URI references: link rot and content drift. Link rot represents the case where a URI ceases to exist and hence the request typically returns a “404 – Page not found” response. Content drift describes the case where the resource identified by its URI changes over time and hence, as time goes by, the request returns content that becomes less and less representative of what was originally referenced.

Our recent [study](#), part of the research track of the [Hiberlink project](#), investigates the extent to which science, technology, and medicine (STM) articles are subject to reference rot. Numerous link rot studies for the general web as well as for scholarly communication have been published over the years but the investigation into the aspect of content drift as well as the sheer scale of our experiments make our study unique to date. We assembled three corpora (arXiv, Elsevier, PubMed Central) consisting of more than 3.5 million scholarly articles published between 1997 and 2012. Out of the 4 million URIs that we found (in about 1.8 million of the considered articles), more than 1 million became subject to our analysis as they referenced web at large resources (blogs, project websites, etc.) as described above. URIs that referenced STM articles were excluded from our study on the assumption that existing persistent identifier and archival infrastructure makes such links immune to reference rot.

Our results show an alarming link rot ratio for all three corpora: 13% of arXiv, 22% of Elsevier, and 14% of PMC articles published in 2012 suffer from link rot. These numbers only increase for older articles, for example, for articles published in 2005 the corresponding numbers are 18%, 41%, and 36%.



The graphic shows links from the three STM corpora that were studied (right hand side) to the six top level domains they link to (left hand side). The colored portions shows links that are healthy, the grey portions – by far the largest – show links that are infected by reference rot.

We investigated the notion of content drift by checking for archival copies (Mementos) of the referenced web at large resource in web archives. We considered a Memento created within 14 days of the article’s publication date as representative to what was originally intended for the reference. While this time window is chosen somewhat arbitrarily, given the dynamic character of the web, we make the case that chances are fairly high that the referenced web at large resource has changed within 14 days and hence does not necessarily represent the intended content of the reference anymore. Throughout all three corpora, about 75% of all URI references lack such representative Mementos.

To elevate these numbers to the article level, we defined an article as being subject to reference rot if it contains at least one URI reference that is either subject to link rot or is lacking a representative Memento. We extrapolated our Elsevier numbers to the scale of the overall STM article landscape and arrived at eye-opening numbers: about 20% of STM articles published in 2012 suffer from reference rot. The remaining 80% either do not contain any URI references or, when they do, they all still work and a representative Memento is available for each. When considering only those articles that contain one or more URI references, the ratio of infected articles published in

2012 increases to 70%.

Reference rot is clearly a significant problem, and, in addition to quantifying its extent, the Hiberlink project also explores ways to mitigate it. The typical strategy to address the problem, followed by those who actually care about it, is as follows:

- When linking to a web page, a snapshot of the state of the page at linking time is created in a web archive. Several web archives provide on-demand snapshot functionality, including the [Internet Archive](#), [archive.today](#), and [perma.cc](#).
- With the snapshot created, rather than linking to the original web page, a link to the snapshot is put in place.

For example, we are writing this post on January 21 2015. And we want to link to <http://www.w3.org/>. That's the W3C's page, and it changes rather frequently. In order for future readers of this blog post to see the same W3C content that we saw when linking, we create a snapshot, say in the [archive.today](#) web archive. That snapshot is available at <https://archive.today/r7cov> and, rather than linking to <http://www.w3.org/>, we link to <https://archive.today/r7cov>.

While the creation of the snapshot is definitely an essential step in the right direction, there are problems with the linking approach:

- Linking to the snapshot <https://archive.today/r7cov>, assumes that the [archive.today](#) web archive will exist forever. Unfortunately, there are plenty of indications that web archives do not have eternal life either. If the web archive in which we created the snapshot suffers a temporary glitch, moves its content to another web location, or ceases to exist, visiting the snapshot becomes impossible.
- When linking to the snapshot <https://archive.today/r7cov>, the URI of the W3C's page, <http://www.w3.org/>, is lost. As a result, future readers of this post cannot visit the W3C page to see its evolved state.

Link decoration is a way to address these problems and to increase the chances that links will lead to meaningful content, even a long time after they were put in place. In order to maximize link robustness, the following information should be available, in a machine-actionable manner, for a link:

- the URI of the snapshot, in our example <https://archive.today/r7cov>
- the URI of the original resource, in our example <http://www.w3.org/>
- the datetime of linking, in our example January 21 2015.

The latter two information elements can be used to automatically find snapshots in other web archives in case [archive.today](#)'s service is interrupted, and the snapshot <https://archive.today/r7cov> becomes inaccessible as a result.

Discussions with interested parties are still underway regarding the best way to convey this information on a link. Until further notice, and for demonstrations purposes, the information is conveyed using [HTML5's attribute extensibility mechanism](#). Using that approach, [this robust link to the W3C home page](#) looks as follows:

```
<a href="http://www.w3.org/"  
data-versionurl="https://archive.today/r7cov"  
data-versiondate="2015-01-21">this robust link to the W3C home page</a>
```

The [Memento Time Travel extension for Chrome](#) makes these link decorations accessible when right-clicking on the link. Try it with [a version of the reference list of our above mentioned article](#) in which links to web at large resource were decorated. For more information on link decoration, check out the [Robust Links](#) site.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the

London School of Economics. Please review our [Comments Policy](#) if you have any concerns on posting a comment below.

About the Authors

Martin Klein is an Analyst at the University of California Los Angeles Research Library. He can be found on Twitter @mart1nkle1n

Herbert Van de Sompel is the team leader of the Prototyping Team at the Research Library of the Los Alamos National Laboratory. The Team does research regarding various aspects of scholarly communication in the digital age. He can be found on Twitter @hvdsomp

- Copyright © The Author (or The Authors) - Unless otherwise stated, this work is licensed under a Creative Commons Attribution Unported 3.0 License.