Content referenced in scholarly articles is drifting, with negative effects on the integrity of the scholarly record

2/23/2017

In their 2015 post, ***Martin Klein*** *and* ***Herbert Van de Sompel*** *reported on the beginnings of an investigation into 'reference rot' in scholarly articles. This term incorporated 'link rot', whereby referenced web-at-large resources vanished from the web altogether, and 'content drift', whereby a resource's content changed over time to such an extent as to cease to be representative of that originally referenced. Results from the initial study found that between 13% and 22% of references suffered from link rot. Here, Klein and Van de Sompel describe the findings of a more recent study assessing content drift. Results show as much as 75% of referenced content had changed to some degree in just three years, raising significant concerns over the integrity of the scholarly record. However, increased adoption of 'robust links' offers a viable solution to this problem.*

More and more scholarly articles not only contain references to books and other scholarly publications but also to 'web-at-large' resources such as project websites, blogs, videos, presentations, ontologies, and source code repositories. These resources are commonly referenced by means of their HTTP URI (uniform resource identifier) and, since they live on the web, are subject to the same dynamics as all other web resources. In particular, they suffer from link rot and content drift.

Link rot represents the case where the resource identified by a URI vanishes from the web. As a result, a URI reference to the resource ceases to provide access to referenced content. Content drift describes the case where the resource identified by a URI changes over time. The resource's content evolves and can change to such an extent that it ceases to be representative of the content originally referenced. We coined the term 'reference rot' to denote the combination of the two problems. Inarguably, reference rot is of major detriment to the integrity of the web-based scholarly record. A reader who visits a web-at-large resource by following a URI reference in an article, some time after its publication, is led to believe that the resource's content is representative of what the author originally referenced. However, due to reference rot, this may very well not be the case. In case of link rot, it will be obvious to the reader that something is awry because she will receive a "404 Not Found" message. In the case of content drift, the reader has no ability to determine whether the obtained content is the same as when it was originally referenced by the author.

Image credit: Drift by Moyan Brenn. This work is licensed under aCC BY 2.0 license.

In a previous post we reported on our 2014 study in which we began to investigate the notion of reference rot in scholarly articles. One of the aspects that set this study apart was its unprecedented scale, with a corpus of more than 3.5 million articles from which we extracted more than one million URIs to web-at-large resources. By checking all URIs on the live web, we were able to precisely quantify link rot. We found that between 13% and 22% of references suffered from it less than two years after publication of the referencing article. Factoring in the numbers for content drift, we arrived at the conclusion that about 20% of recently published scholarly articles (one in five) suffer from reference rot.

In our most recently published study we applied a more sophisticated and accurate approach to assess the notion of content drift for referenced web-at-large resources. Our methodology was based on finding an archived copy (a Memento) of a URI reference that is representative of what it was at the time the referencing article was published. Once a representative Memento was discovered, we compared it to its counterpart on the live web, if it still existed, and assessed whether the content had drifted.

Using the Memento Framework and 19 public web archives, we obtained the Memento created closest to and also prior to the article's publication date (Memento_Pre) as well as the Memento created closest to and also after that date (Memento_Post). If these two Mementos were identical, we declared them representative of the reference as it was at article publication time. Judging whether two web resources are identical can be done in different ways. One feasible approach is to measure the similarity of the textual content of the web page. We chose to apply four different text similarity measures to assess the similarity of the Memento_Pre/Post pair that was created around the publication date of the referencing article. The measures we used are Simhash, Jaccard, Sørensen-Dice, and Cosine. Since their underlying algorithms are focused on different aspects of textual similarity, we are confident that the combination of their scores provides a meaningful assessment of similarity. In addition, we opted for a very stringent filter; only if all four measures returned a perfect similarity score when comparing our two Mementos did we consider them identical and hence representative. We applied this methodology to each of our more than one million URI references of interest and found that 64% have Memento_Pre/Post pairs. Of these, 35% have representative Mementos.

This fraction of URI references for which we found a representative Memento was then subject to our content drift investigation. We selected one of the Mementos from the previous comparison (it is irrelevant which one as they are identical) and compared it to the version of the URI reference on the live web using the same content similarity analysis as seen before.

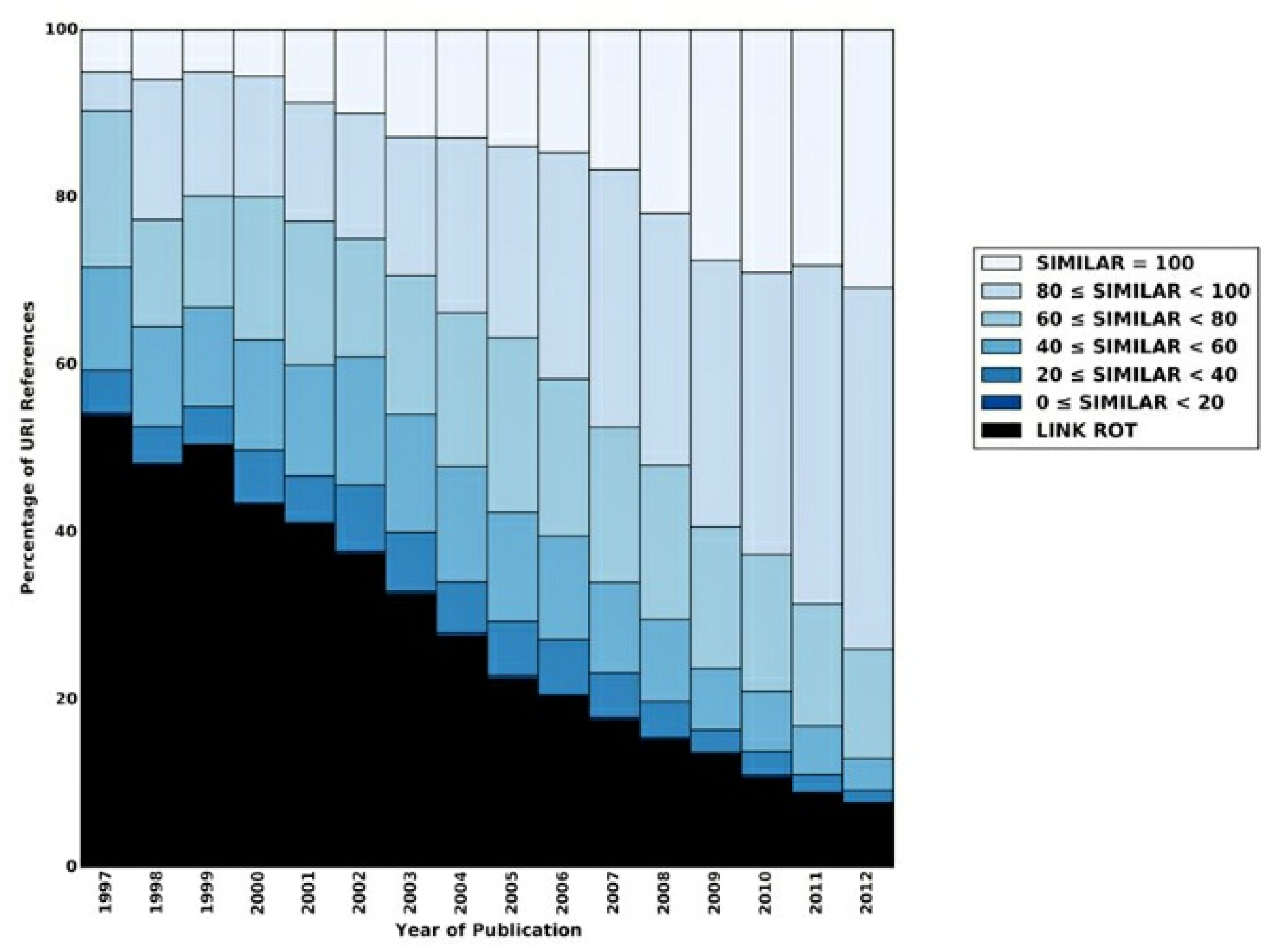

Figure 1: Similarity Ranges for Representative Mementos and Live Web Content per Publication Year—arXiv corpus. This figure was originally published in the article 'Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content' and is published under a CC0 1.0 license.

Figure 1 (above) shows the results of that comparison for URI references extracted from articles in the arXiv corpus published between 1997 and 2012. The colours correspond with ranges of the computed similarity where the lightest blue indicates the maximal similarity. As similarity decreases, the blue gets darker. URI references that are subject to link rot are shown in black. The publication year of the referencing article is shown on the x-axis, and the percentage of URI references is shown on the y-axis. Several alarming observations can be made from this graph. First, the large percentage of link rot confirms our 2014 findings. Second, the content of only one out of four URI references from most recently published articles has not drifted. That means, for papers published in 2012, 75% of referenced content had changed by the time of our study in August 2015. This ratio gradually gets worse for URI references from articles published further in the past. The numbers are very similar for the other two corpora in our study.

The reliable quantifications for link rot and content drift provided in our two studies prove that reference rot is a significant problem in scholarly communication. However, it is one that can be ameliorated. We introduced the concept of Robust Links that, by using Link Decoration, makes links more robust over time and hence helps in combating reference rot. Robust Links are based on a two-step process:

1. Proactively creating an archival snapshot of the referenced web-at-large resource. Web archives such as the Internet Archive, cc, and weblock.io allow a user to do this.
2. Appropriately referencing these snapshots in scholarly literature by using the Link Decoration approach. This approach enhances the reference of a URI with the URI of the archived snapshot and the datetime of archiving, both obtained in step one. Together, these three bits of information allow the seamless navigation from a URI reference to live web content and a representative Memento in a manner that provides appropriate guarantees that the reference remains robust over time.

An example of Robust Links in action can be found in this 2015 D-Lib Magazine article. For more information about Robust Links, see http://robustlinks.mementoweb.org.

In summary, scholarly articles increasingly contain URI references to web-at-large resources that, as they live on the web, are subject to reference rot (link rot and content drift). This has a negative impact on the integrity of the scholarly record. Unlike for scholarly articles, custodians of such web-at-large resources are typically not overly concerned about long-term access to their own content, let alone about the longevity of the scholarly record. As such, reference rot is a problem with roots largely outside the scholarly communication community but one that will have to be solved by that very community. We suggest Robust Links as a possible solution.

This blog post is based on the article 'Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content' by Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin and Claire Grover, published in PLoS ONE (DOI: 10.1371/journal.pone.0167475).

This work originated in the Hiberlink project, a collaborative effort between LANL, EDINA, and the Language Technology Group at the University of Edinburgh.

Note: This article gives the views of the authors, and not the position of the LSE Impact 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 a Scientist in the Prototyping Team at the Research Library of the Los Alamos National Laboratory . 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*