Preparing, Not Repairing

Leah Cannon[1], Brianne Dosch[2], Hannah Gunderman[12]

[1]School of Information Sciences -University of Tennessee, Knoxville
[2]Center for Information and Communication Studies -University of Tennessee, Knoxville

Abstract: LibGuides are widely used in academic libraries to organize information and resources to better serve their users, yet the maintenance of LibGuides across many different academic libraries ranges from neglectful to fastidious, making it difficult to determine the accuracy and helpfulness of information on LibGuides (Randtke & Burrell, 2012). A Scrapy-based Python script developed by the SIS Programming Team was used to crawl the LibGuides of each academic library member of the ARL, and a second script was used to test the HTTP status of each link. Maintaining link validity in LibGuides is a proactive and strong first step toward patron satisfaction, and using a combination of scripts and manual processes can be a proactive way to monitor link validity and fight link rot.
Keywords: libguides, link rot, content drift, dataone, python, link validity

1. Introduction

One of the hallmark tenets of libraries is to provide patrons with consistent and accurate access to information. Resources such as LibGuides are a medium through which to provide patrons with domain-specific information. However, what happens when the LibGuides contain broken or inaccurate URLs, potentially eliminating or complicating access to information? In this paper, we explore the impact of link rot and content drift (holistically referred to as "reference rot") on LibGuides within the websites of members of the Association of Research Libraries (ARL), and introduce a computational technique for automating and expediting a historically-manual link checking process.

In the fall of 2018, DataONE commissioned a white paper accessing the reach of their organization in the websites of members of the ARL. At that time, this research was focused on DataONE's reach regardless of the status of the links to the organization's website. After that project was completed, a follow-up project was developed to test all links to DataONE.org found in the LibGuides at ARL member academic libraries. This follow-up project introduced the use

Received: 22.7.2019 Accepted: 26.2.2020 ISSN 2241-1925

of computational techniques to automate the manual link retrieval processes used in the initial project.

2. Literature

Scholars and librarians spend countless hours developing and curating online tools and resources relevant to their research and their intended users. But when an individual leaves that project or institution, do those tools and resources persist? Can they persist without maintenance? The short answer: no. No scholarly efforts, no matter how cutting edge or wide spread, will persist on the web without strategic and deliberate maintenance efforts to fight what has been coined by the HiberLink Project as "reference rot" (Hiberlink, n.d.).

Klein et al. (2014) define reference rot as "the combination of two problems" faced by online content, which are "link rot: the resource identified...may cease to exist" and "content drift: the resource identified...may change over time...that it ceases to be representative" of the original referenced content" (p. 2). This problem of reference rot has been pursued and studied in many contexts (Fielding, et al., 1999; Donlan, 2007; Isfandyari-Moghaddam, 2011; Zhou, et al., 2014; Gordon-Murnane, 2018) as this is a growing concern as scholarly communication is becoming more and more dependent on reliable and persistent web-content (Burnhill, et al., 2015, p.56). Despite this growing research subject and a breadth of well-known best practices for creating durable web-content, one in five scholarly articles suffer from reference rot (Klein, et al., 2014, Gordon-Murnane, 2018). Therefore, whose responsibility is it to implement said best practices and start fighting link rot and content drift in the scholarly communication landscape?

Dowling (2001) placed the onus of fighting link rot and content drift on the content provider, delineating a 24-month plan for moving web sites. Dowling stated "not many users will take the time and energy to hunt down your library's new URL when their old bookmark fails them" and "even if they did, they wouldn't always know where and how to look" (p.36). Certainly, the content provider plays a role (including DataONE, which this paper will later illustrate), but scholars and librarians that are content creators (and therefore managers) should not be passive.

A great example of how reference rot can negatively affect the efforts of scholars and librarians are the doomed lifecycles of LibGuides. LibGuides are widely used in academic libraries to organize information and resources to better serve their users. Despite this valuable utility, the maintenance of LibGuides across many different academic libraries ranges from neglectful to fastidious, making it difficult to determine the accuracy and helpfulness of information on LibGuides (Randtke & Burrell, 2012).

Jurewicz and Cutler (2013) noted "as websites have grown in complexity, the time-consuming process of coding HTML pages has become a huge task" (p. 76). Fifteen years later, even though current technologies have made the task easier, the number of web sites is growing and increase by millions every year (Netcraft, 2019). LibGuides, being among those millions of sites, are also increasing in complexity and frequency. Therefore, it is becoming increasingly important for libraries to follow the best practices of proactively maintaining LibGuides rather than a reactive position, or just abandoning them to reference rot and disuse.

This research highlights a case study of a successful computational link checking process, introducing a more efficient method than manually checking the status of links. More broadly, this research emboldens LIS professionals to look closely at reference rot in action, not only to find warning signs of eventual link rot (like content drift) but to identify the best practices scholarly content creators (and therefore managers) should be following if a meaningful digital impact will persist beyond original project and research personnel.

3. Methodology

For the initial project, the August 2018 member list of the ARL was used to delineate which LibGuides would be searched for references to DataONE. This member list included 116 academic libraries. Each academic library's web site was manually searched for a mention of the phrase "dataone." If no mention was found, the next step was to look for reference, subject, or LibGuides within the library's site and perform the same search (in several cases, the search for the LibGuides was separate from the search on the main library's page). The final step was to search Google.com using the statement:
"dataone" site:<library or subject guide web site>

If no results were found using the above three methods, the assumption was made that no mention of DataONE was present on the library's web site.

If a mention was found, the URL containing the mention, the page's subject, the party responsible for the page's content (if that party could be identified), and the DataONE page linked were all recorded. Links were not tested for validity at this time.

Heylighen (2013) describes human computation as a process of letting people "perform those tasks that are too difficult for a computer program, while using computers to do the tasks that are difficult or tedious for people" (p. 899). While the manual method of link retrieval in the initial project described above produced extensive results, the amount of time and tediousness necessary for locating and extracting the links negated the overall idea of efficiency within link maintenance. Accordingly, the University of Tennessee-Knoxville's School

of Information Sciences Programming Research Group[1] (hereafter SIS-PRG) endeavored to automate this link checking process to allow for quicker identification and repair of link rot. A Scrapy-based Python script was adapted[2] to automate the link checking process to locate if and where certain databases are referenced within an academic research library website. [Image 1]. Scrapy is an "application framework for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like data mining, information processing or historical archival" (Scrapy at a glance, 2011). Scrapy allows for the development of *spiders*, scripts which parse through a website, clicking links and crawling through each individual page. This script, named ARLSpider, is a web scraping script which looks for URLs to a defined domain within a website, following links until the entire website is scraped and reporting the results to a .csv file. Since the DataONE.org links had already been pulled manually, the data from was used by the SIS-PRG to validate the links pulled by ARLSpider.

```
class ArlSpider(CrawlSpider):
    name = 'arlspider'

    # All links pointing to these domains are followed by the spider
    domains_for_scraping = ['libguides.utk.edu/']
    # All links pointing to these domains are added to the CSV file
    domains_to_look_for = ['dataone.org']
    # The list of URLs to start on
    start_urls = ['https://libguides.utk.edu/']

    def start_requests(self):
        for url in self.start_urls:
            yield scrapy.Request(url, callback=self.parse_page, dont_filter=False)

    def parse_page(self, response):
        """Parse a page"""

        # This method can yield multiple links and requests
        # If a request is yielded, then the link is followed
        # When a link is yielded, then the link will be added to the CSV file

        # Extract unique links on the current page
```

Figure 1: Portion of ARLSpider Python Script.

Once the SIS-PRG had thoroughly debugged and tested the script, it was run for the LibGuides at each academic library member in the ARL. Some domains did not allow scraping; webcrawlers such as ARLSpider can retrieve information from websites at a higher and quicker volume than manual methods performed by humans, which can have a detrimental effect on some websites. Therefore,

[1] The SIS Programming Research Group was established in Fall 2018 as a research collective dedicated to providing solutions for automating information retrieval tasks

[2] With special thanks to the original developer of the script upon which ARLSpider was adapted, Kevin Jacobs, at https://www.data-blogger.com/2016/08/18/scraping-a-website-with-python-scrapy/

some websites with under-powered servers will restrict the presence of webcrawlers. In the case of these particular websites, even if links to DataONE.org were found using the previous manual method, they were not included with the scraped data in order to maintain methodological integrity.
All URLs found through ARLSpider were then tested using a second Python script, which returned the HTTP status for each link. If a link redirected to another link, the script captured the number of redirects, each page in the redirect chain, and the final URL.

4. Results

4.1. Initial examination

In the initial, manual process, 78 of the 116 ARL member academic library members mentioned DataONE at least once, 38 had no mention. Of the 78 libraries' web sites that mentioned DataONE, there was an average of 4.5 mentions per library (if academic libraries with no "dataone" mention are included, the average becomes 3 mentions per library). A total of 357 mentions were found in the manual exploration, with nineteen that either had either a broken link or no link to DataONE.

A subject was recorded for each page that contained a mention of DataONE. If the subject was not identified or easily determined, it was recorded as "Research Data." Pages identified as "Research Data" linked to DataONE most frequently, with 206 links (57.7% of all links). The most frequently linked-to site on DataONE's website was Best Practices with 108 linking pages, followed by 72 pages with links to the main DataONE site. Data Management Planning was third most frequently linked with 27 links. Of the 108 links to the Best Practices directory, only 34 pages at 17 libraries actually linked to the top level at https://www.dataone.org/best-practices.

4.2. Further examination

In December 2018, the manually collected data was presented to DataONE in a white paper, the focus of which was the reach of DataONE among ARL academic libraries. At that time, the team decided to take a look at the validity of the links rather than the content. Concurrently, the computational method was being tested by the SIS-PRG. A project was undertaken to retrieve all DataONE.org links from ARL member academic libraries, and to test those links once collected.

A total of 371 links to DataONE.org were found using the computational method, with links to 66 URLs on the DataONE.org domain [Table 1].

HTTP code	Count of URL	Distinct Count of URL
200	242	43
301	127	21

404	2	2
Total	371	66

Table 1: URL counts by HTTP code

Sixty-six, or 57%, of the 116 academic library members of ARL link to a site on the DataONE.org domain [Table 2].

University	Count of URL	Distinct Count of URL
Texas A&M University	36	1
Virginia Polytechnic Institute and State University	32	2
Purdue University	24	5
Yale University	18	9
University of California, Santa Barbara	17	7
Cornell University	15	7
Tulane University	15	4
University of Oklahoma	13	9
Rutgers University	12	10
University of Florida	10	5
University of Tennessee, Knoxville	8	6
University of Hawaii at Manoa	8	4
Georgia Institute of Technology	8	2
Florida State University	7	4
University of Michigan	7	6
Boston College	7	3
Stony Brook University	6	5
Oklahoma State University	6	4
University of Illinois at Urbana-Champaign	6	3
Washington State University	5	5
Georgetown University	5	4
University of Missouri	5	3
University of Iowa	4	2
Auburn University	4	3
University of Miami	4	2
University of California, Davis	4	2
George Washington University	4	4
University of California, Irvine	4	3

University of Texas	3	3
Pennsylvania State University	3	3
Virginia Commonwealth University	3	3
Queens University	3	3
University of South Carolina	3	3
Syracuse University	3	2
University of Utah	3	2
University of Manitoba	3	1
Western University	3	2
Ohio State University	3	1
Kansas University	3	2
University of North Carolina	3	2
University of Illinois at Chicago	2	2
Northwestern University	2	1
University of Washington	2	2
University of Calgary	2	2
University of British Columbia	2	1
University of Ottawa	2	1
University of Massachusetts	2	1
University of Pennsylvania	2	1
University of Wisconsin-Madison	2	1
University of Saskatchewan	2	2
New York University	2	2
University of Kentucky	2	1
McGill University	2	2
Iowa State University	2	2
Arizona State University	2	2
Temple University	1	1
University of Oregon	1	1
University of Georgia	1	1
University of New Mexico	1	1
Kent State University	1	1
University of California, Berkeley	1	1
University of Pittsburgh	1	1
University of Notre Dame	1	1
University of Nebraska-Lincoln	1	1
University of Chicago	1	1
University of Colorado Boulder	1	1

Total	**371**	**66**

Table 2: URL Counts by University

Turning attention to the HTTP status of the links, the majority (242 out of 371, or 65% of links) returned an HTTP status of 200, OK, or as defined by Fielding, et al, "the request has succeeded" (1999). Only two (0.5%) returned 404, or "The server has not found anything matching the Request-URI" (Fielding, et al, 1999). Slightly over a third, 34% or 127 of links, returned a 301 status, Moved Permanently, or "The requested resource has been assigned a new permanent URI and any future references to this resource SHOULD use one of the returned URIs" (Fielding, et al, 1999).

Taking a closer look at the URLs with a 301 status, http://dataone.org was linked by eleven library ARL members. This link redirects twice ending at https://www.dataone.org. The first redirect adds the secure "HTTPS" prefix to the link, and the second adds www. to the start. A review of the remaining twenty distinct links (116 total occurrences of these links) that redirect reveal that the only difference in the referring (the link specified by the LibGuide) and the landing URL (where the redirect eventually "lands") is the addition of the HTTPS prefix. Although the landing URL is the same as the referring URL, Fielding, et al. (1999), note in the description of 301 status, "clients with link editing capabilities ought to automatically re-link references to the Request-URI to one or more of the new references returned by the server, where possible."

5. Conclusions and Future

With this research, we not only seek to highlight issues of link rot within academic library LibGuides, but we hope to increase the use of computational techniques within library and information science environments. Automating manual tasks such as link checking not only ensures patrons are receiving consistent access to information, but also frees up time for library professionals to pursue other endeavors.

Maintaining link validity in LibGuides is a proactive and strong first step toward patron satisfaction, and using a combination of scripts and manual processes can be a proactive way to monitor link validity and fight link rot. As Raub and Liao (2012) state, "the uncertainty and volatility involved in customer service require frontline employees to take personal initiative to anticipate customer needs, prevent and remove potential obstacles in service delivery before a problem 'hits the surface'" (p. 655). Schloss and Ravel (2018) found that "rapid advances in sequencing technology, data curation, databases, and statistical techniques present an additional threat to reproducibility because resources and what are considered best practices are constantly evolving," and "unfortunately, the developer of the web resources must ensure that the resource remains active." Libraries currently enlist both manual and computational methods for

link checking, with varying degrees of success with the latter method. For example, after finding that their "automated link checker...tool does not work well for identifying database-level access issues," Mortimore and Minihan (2018) at Georgia Southern University reported that "Every quarter, a paraprofessional or student worker clicks through all of our curated book and link assets and notes any authentication or access issues in a dedicated Google sheet" (p. 7) (a process that Mortimore and Minihan admit takes half a day) and afterwards "the Discovery Services Librarian reviews the access notes and troubleshoots any problems" (p. 7). Certainly, this manual process is thorough, but the process is time consuming and at least two individuals are involved in the process. Mortimore and Minihan specifically mention LibGuides' built in link checker, but note it "does not work well for identifying database-level access issues" (p. 7). In addition, Randtke and Burrell found that this link checker will only check links in bulleted lists, and will not check links in text paragraphs, further stating that "libraries need to find other avenues to identify broken links" (2012). This is an important shortcoming to acknowledge when discussing link scraping and testing Python scripts as our research does.

As an additional finding, a portion of the links collected referenced DataONE's collection of research data management guides and resources. In an information environment seeking increasing access to open data and open science (Monastersky, 2013), libraries and research institutions have a responsibility to provide comprehensive research data management services to patrons and researchers. Many libraries are increasingly investing in RDM resources, whether through the hiring of librarians and consultants specifically targeted at RDM services, or by providing workshops and research guides aimed at RDM. The presence of links to DataONE's RDM materials signifies an investment in RDM infrastructure among academic research libraries.

References

Burnhill, P., Mewissen, M., & Wincewicz, R. (2015). Reference Rot in Scholarly Statement: Threat and Remedy. *Insights 28*(2), 55-61.

Donlan, R. (2007). Boulevard of broken links: Keeping users connected to e-journal content. *The Reference Librarian, 48*(1), 99-104.

Dowling, T. (2001). One Step at a Time. *Library Journal*, *126*(17), 36. Retrieved from http://search.ebscohost.com.proxy.lib.utk.edu:90/login.aspx?direct=true&db=eue&AN=502881179&scope=site

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T. (1999, June). *Hypertext Transfer Protocol -- HTTP/1.1: 10 Status Code Definitions*. Retrieved from https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

Gordon-Murnane, L. (2018). Linkrot content drift = reference rot. *Online Searcher, 42*(6), 10-17.

Heylighen, F. (2013). From human computation to the global brain: The self-organization of distributed intelligence. In P. Michelucci (Ed.). *Handbook of human computation* (pp. 897-910). Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-1-4614-8806-4.pdf

Hiberlink (n.d.) "About," retrieved from http://hiberlink.org/about.html

Isfandyari-Moghaddam, A., & Saberi, M-K. (2011). The life and death of URLs: The case of Journal of the Medical Library Association. *Library Philosophy and Practice*, 1. Retrieved from http://digitalcommons.unl.edu/libphilprac/592/.

Jurewicz, L., & Cutler, T. (2003). *High Tech, High Touch: Library Customer Service Through Technology*. Chicago: American Library Association.

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, *9*(12), 1–39. Retrieved from https://doi.org/10.1371/journal.pone.0115253

Monastersky, R. (2013). The library reboot: as scientific publishing moves to embrace open data, libraries and researchers are trying to keep up. *Nature*, *495*(7442), 430+. Retrieved from http://link.galegroup.com.proxy.lib.utk.edu:90/apps/doc/A324206492/AONE?u=tel_a_utl&sid=AONE&xid=10d07d9c.

Mortimore, J., & Minihan, J. (2018). Essential audits for proactive electronic resources troubleshooting and support. *Library Hi Tech News 35*(1), pp. 6-10. Retrieved from https://www-emeraldinsight-com.proxy.lib.utk.edu/doi/full/10.1108/LHTN-11-2017-0085

Netcraft. (2019). January 2019 web server survey. Retrieved from https://news.netcraft.com/archives/2019/01/24/january-2019-web-server-survey.html

Randtke, W., & Burrell, M. (2012, June 1). Tools for Reducing and Managing Link Rot in LibGuides. *Code4Lib Journal*, (17). Retrieved from https://doaj.org/article/82b031449996477aa4d462d29b37600c

Raub S and Liao H (2012) Doing the right thing without being told: Joint effects of initiative climate and general self-efficacy on employee proactive customer service performance. Journal of Applied Psychology 97(3): 651–667.

Schloss, P., & Ravel, J. (2018). Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio*, *9*(3). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5989067/.

Scrapy at a Glance. (2011). Retrieved from https://doc.scrapy.org/en/0.14/intro/overview.html

Sharp, C. (2017). Social media and other research tools. *American Bar Association*. Retrieved from https://www.americanbar.org/groups/gpsolo/publications/gpsolo_ereport/2013/december_2013/social_media_and_other_research_tools/.

Sicilia, M.A., Garcia-Barriocanal, E., Sanchez-Alonso, S., Cuadrado, J.J. (2019). Decentralized persistent identifiers: A basic model for immutable handlers. *146*, 123-130.

Zhou, K, Tobin, R and Grover, C (2014). Extraction and Analysis of Referenced Web Links in Large-Scale Scholarly Articles. Proceeding of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'14), 451–452.