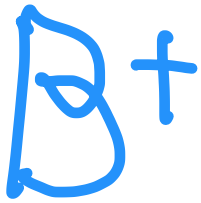
Reference Rot in the Repository: A Case Study of Electronic Theses and Dissertations (ETDs) in an Academic Library



Mia Massicotte and
Kathleen Botter

ABSTRACT

This study examines ETDs deposited during the period 2011-2015 in an institutional repository, to determine the degree to which the documents suffer from reference rot, that is, linkrot plus content drift. The authors converted and examined 664 doctoral dissertations in total, extracting 11,437 links, finding overall that 77% of links were active, and 23% exhibited linkrot. A stratified random sample of 49 ETDs was performed which produced 990 active links, which were then checked for content drift based on mementos found in the Wayback Machine. Mementos were found for 77% of links, and approximately half of these, 492 of 990, exhibited content drift. The results serve to emphasize not only the necessity of broader awareness of this problem, but also to stimulate action on the preservation front.

INTRODUCTION

A significant proportion of material in institutional repositories is comprised of electronic theses and dissertations (ETDs), providing academic librarians with a rich testbed for deepening our understanding of new paradigms in scholarly publishing and their implications for long-term digital preservation. While academic libraries have long collected and preserved hard copy theses and dissertations of the parent institution, the shift to mandatory electronic deposit of this material has conferred new obligations and curatorial functions not previously incorporated into library workflows. By highlighting ETDs as a susceptible collection deserving of specific preservation actions, we draw attention to some unique responsibilities for libraries housing university-produced content, particularly as scholarly information continues its shift away from commercial production and distribution channels. As Teper and Kraemer point out in their discussion of ETD program goals, "without preservation, long-term access is impossible; without long-term access, preservation is meaningless."[1]

What Is Reference Rot, And Why Study It?

In addition to *linkrot* (where a link sends the user to a webpage which is no longer available),

Mia Massicotte (Mia.Massicotte@concordia.ca) is Systems Librarian, Concordia University Library, Montreal, Quebec, Canada. **Kathleen Botter** (Kathleen.Botter@concordia.ca) is Systems Librarian, Concordia University Library, Montreal, Quebec, Canada.

there are webpages that remain available, but whose contents have undergone change over time--known as *content drift*. This dual phenomena of linkrot plus content drift has been characterized as *reference rot* by the Hiberlink project team,[2] and has important implications for digital preservation. Since theses and dissertations are original works born digital by virtue of mandatory deposit programs, a university's ETD program is effectively a digital publishing initiative, accompanied by a new universe of responsibility for its digital preservation.

Due to the specialized nature of graduate-level research, ETDs frequently include links to resources on the open web, for example, personal blogs, project websites, and commercial entities. Digital Object Identifiers (DOIs), useful in the context of published literature, do not apply to URLs on the free web, which are DOI-indifferent. Open web links also fall outside the scope of preservation initiatives such as LOCKSS (Lots of Copies Keep Stuff Safe)[3] which aim to safeguard the published literature. With increasing frequency, researchers are citing newer forms of scholarship, which do not readily fall under the rubric of published literature. Moreover, since thesis preparation is conducted over a period of time typically measured in years, links cited therein are likely to be more vulnerable to linkrot and content drift by the time of manuscript submission.

Yet despite the surfeit of anecdotal daily evidence that URLs vanish and result in dead links, Phillips, Alemneh, and Ayala point out that "by and large academic libraries are not capturing and maintaining collections of web resources that provide context and historical reference points to the modern theses and dissertations held in their collections."[4] Since an ETD comprises a unique form of scholarly output produced by universities, and simultaneously satisfies the parent institution's degree-granting apparatus, as well as reflecting its academic stature on the international stage, the presence of reference rot in this body of literature is of particular concern and worthy of immediate attention.

Smoking Guns

There has been no shortage of evidence reporting on the linkrot phenomena over the last two decades. Koehler, whose initial study on linkrot appeared in *JASIS* in 1999, periodically revisited, analyzed, and reported on the same set of 360 URLs collected in his original study.[5,6,7] In 2015, upon the twenty-year benchmark of the original data collection, Oguz and Koehler reported in *JASIS* that only 2 of the original links remained active.[8]

A number of foundational studies, including Casserly and Bird,[9] Spinellis,[10] Sellitto,[11] Falagas, Karveli, and Tritsaroli,[12] and Wagner et al.[13] have reported on linkrot occurring in professional literature. Sanderson, Phillips, and Van de Sompel provide a table of 17 well-known linkrot studies, comparing overall benchmarks, and supplying a succinct summary of the scope of each study.[14] Linkrot also gained further important exposure with the Harvard Law School study by Zittrain, Albert, and Lessig, which found that 70% of 3 Harvard law journal references, and 49.9% of URLs in Supreme Court opinions examined, no longer pointed to their originally cited sources.[15]

Members of the Hiberlink project, which set out to examine "a vast corpus of online scholarly publication in order to assess what links still work as intended and what web content has been successfully archived using text mining and information extracting tools" have been pivotal in making the case for reference rot.[16] Hiberlink demonstrated that failure to link to cited sources was due not only to linkrot, but also to web page content which changed over time.[17]

A new dimension of the digital preservation universe was thrown into sharp relief with follow-up study by Klein et al. (2014), which examined one million web references extracted from 3.5 million Science, Technology, and Medicine (STM) articles published in Elsevier, PubMed Central, and ArXiv, between the years 1997 and 2012. The study concluded that one in five articles suffers from reference rot.[18] Though the study focused on STM articles, its authors drew attention to theses and dissertations as a susceptible class of material. Analyzing the same set of links extracted from this large STM corpus, Jones et al. (2016) recently reported that 75% of referenced open web pages demonstrated changes in content.[19]

ETDs — A Susceptible Collection

The digital preservation part of institutionally mandated ETD deposit has yet to have its dots fully connected to the rest of the diagram. After four years of research into academic institutions' ETD programs, Halbert, Skinner, and Schultz reported that close to 75% of respondents surveyed had no preservation plan for their ETD collections.[20] Despite the prevalence of linkrot studies, linkrot in ETDs has not been subjected to similar scrutiny, and the implications of disappearance of content is underappreciated. While mandatory deposit programs have become relatively commonplace, focus has largely remained on policy and implementation aspects, metadata quality, interoperability and conformance to standards.[21,22]

There are few studies which focus on institutional repository link content. The study conducted by Sanderson, Philips, and Van de Sompel (2011) was a large-scale examination of two repositories.[23] 400,144 papers deposited in ArXiv, and 3,595 papers in the University of North Texas (UNT) digital library repository were studied, and more than 160,000 URLs examined. Links were analyzed for persistence and the availability of mementos, that is, whether prior versions of the page existed in a public web archive, such as the Internet Archive's Wayback Machine. For 72% of UNT URLs, either mementos were available, or the resource still existed at its original location, or both. Although 54% (9,880) were available in one or more international web archives, 28% (5,073) of UNT's ETD links were found to no longer exist, nor had they been archived by the international archival community.

Phillips, Alemneh, and Ayhala looked at overall general patterns and trends of URL references in repository ETDs, examining 4,335 ETDs between the years 1999-2012 in the UNT repository.[24] The team analyzed 26,683 unique URLs in 2,713 ETDs containing one or more links, finding an overall average of 10.58 unique URLs per ETD with one or more links. The UNT team provided a

breakdown of domain and subdomain occurrence frequency, and indicated areas of future investigation into content-based URL linking patterns of ETDs.

ETD link decay was studied by Sife and Bernard, who performed a citation analysis on URLs in 83 theses published between 2007 and 2011 at Tanzania's Sokoine National Agricultural Library.[25] 15,468 citations were examined, 9.6% (1,487) of which were open web citations. URLs were considered active if found at the original location, or available after a URL redirect. The authors manually tested URLs over a period of seven days to record their accessibility, noting down inaccessible URLs error messages and domains, and analyzing the types of errors encountered. The authors calculated that it took only 2.5 years for half of the web citations to disappear.

At the ETD2014 conference,[26] an important study of 7,500 ETDs in 5 U.S. universities was presented. Of 6,400 ETDs defended between 2003 and 2010, approximately 18% of open web link content was confirmed as lost, and a further 34% at risk of loss, that is, live links which lacked an archived copy.[27] Though the results of that particular study have not been formally published, it was briefly summarized in a session held at the 38th UKSG Annual Conference in Glasgow, Scotland in March 2015, an account of which was subsequently published by Burnhill, Mewissen, and Wincewicz in *Insights*.[28]

Given the scarcity of published literature on link content as found in ETDs, this present study which examines reference rot in ETDs in an academic institutional repository is unique, draws attention to an important digital collection which is vulnerable to loss, and highlights need for action.

BACKGROUND AND CONTEXT

Concordia University is a comprehensive university located in Montreal, with a student population of 43,903 full-time equivalents in 2015, of which 7,835 were graduate students. 27 PhD programs were offered in 2015,[29] and 43 programs at the Masters level. Faculties of Arts and Science, Engineering and Computer Science, Fine Arts, and Business have a thesis requirement, and produce upwards of 350 Masters and 150 PhD dissertations annually. The broad disciplines, and the departmental clusters used in this study are shown in Table 1.

Prior to the thesis deposit mandate, Concordia University Library housed hard copy versions of theses and dissertations in the collection. In 2009, the Library launched Spectrum, Concordia's Eprints institutional repository, playing a leadership role in Spectrum's implementation and policy development, and providing training and support to the School of Graduate Studies regarding submission and management of theses for deposit. Following a successful pilot project, the Graduate Studies Office ceased accepting paper manuscripts, and mandated electronic deposit of all theses and dissertations into Spectrum as of spring 2011.

Discipline	Department	Discipline	Department
Arts	Applied Linguistics Communication Economics Educational Technology History Hist and Phil of Religion Humanities Philosophy Sociology Political Science Psychology Religion	Business*	Decision Sciences and MIS Finance Management Marketing
Engineering**	Building Engineering Civil Engineering Computer Science Comp Sci & Software Eng Electrical and Comp Eng Industrial Engineering Info Systems Security Mechanical Engineering	Fine Arts	Art Education Art History Film and Moving Image Studies Industrial Design Fine Arts Performing Arts
Science	Biology Chemistry Mathematics Physics Exercise Science		

Table 1. Summary of departmental clusters used in this study

* John Molson School of Business
** Engineering & Computer Science

METHODOLOGY

We concentrated on PhD dissertations (henceforth ETDs) in Spectrum in order to limit the scope of the project; Master's theses were excluded. A 5-year period was chosen, beginning with the first semester of mandatory deposit, spring 2011, through fall 2015, a total of 720 ETDs. Since Concordia ETDs are released for publication immediately following convocation, the University's official convocation dates were used to identify the set of documents to be downloaded and examined.

We proceeded in phases: first downloading ETDs from Spectrum and converting to a text format that could be examined for patterns; then extracting links from each and testing programmatically

for linkrot; then drawing a stratified random sample of active URLs and visiting them to determine if content drift had taken place. Our methodology for link extraction was similar to those described by Klein et al.,[30] and Zhou, Tobin, and Grover.[31] During the dissertation download stage, 36 ETDs with embargoed content were encountered and eliminated. ETDs were then converted from existing PDF/A format to xml. A further 20 documents failed to convert due to nonstandard or complex formatting which resulted in unreadable, garbled characters. These documents resisted multiple conversion attempts, and since they could not be mined, had to be eliminated. A final total of 664 ETDs were successfully converted using three different tools: 97% (644) were converted using PDFtoHTML,[32] the remaining 3% by either givemetext (14)[33] or Adobe Acrobat (3).

A spot check of documents was sufficient evidence that many links occurred throughout the text body. Since we intended to extract URLs to the open web, we wanted to err on the side of detecting more links, rather than easily-identifiable well-formed URLs. Links were mined from the body of the text in a manner similar to the study carried out at UNT.[34] We wanted a regular expression which would catch as many URLs as possible, expecting to manually clean the link output before further processing. We tested multiple regular expressions[35] against a small sample of our converted ETDs and compared the results. We selected one which seemed well-suited for our purpose, as it was liberal in detecting links throughout the text, was able to extract links which contained obvious omissions and problems — for example, those that lacked http:// prefixes — but also caught non-obvious errors, such as ellipses in long URLs. We considered how de-duplication of extracted links might affect the outcome, and opted to count each link as an individual instance. Manual cleanup included catching URLs that broke across new lines, identifying false hits such as titles containing colons and DOIs, and adding escape encoding characters for "&" and "%" in order to generate a clean URL for use in the next step of the process.

METHODOLOGY — Linkrot collection

A script programmatically used the cURL command line tool to visit each link and fetch the http response code in return.[36] An output listing was produced for each doctoral dissertation, comprised of the original URLs, the final URLs, and the http response codes. Link output for each of the converted 664 ETDs was collected from December 2015 to January 2016, with the fall 2015 semester checked in March 2016.

76% (504 of 664) of ETDs contained one or more links, the highest number of links (5,946) falling into the Arts group. 24% (160 of 664) of ETDs contained no links. For the 5-year period, the broad discipline breakdown of documents examined, the number of ETDs with links, and the number of links extracted are shown in Table 2. Converted ETDs by publication year, broken out by broad disciplines, are shown in Figure 1.

Discipline	Number of PhD ETDs in Spectrum	ETDs converted*	Contain no links	Contain links	Number of links extracted
Arts	210	195	31	164	5,946
Business	45	43	12	31	210
Engineering	351	326	82	244	3,259
Fine Arts	28	25	2	23	1,728
Science	86	75	33	42	294
Total	**720**	**664**	**160**	**504**	**11,437**

Table 2. 5-year period, 2011-2015, summary of documents examined and links extracted

* 56 documents in total eliminated (36 embargoed; plus 20 which failed to convert).

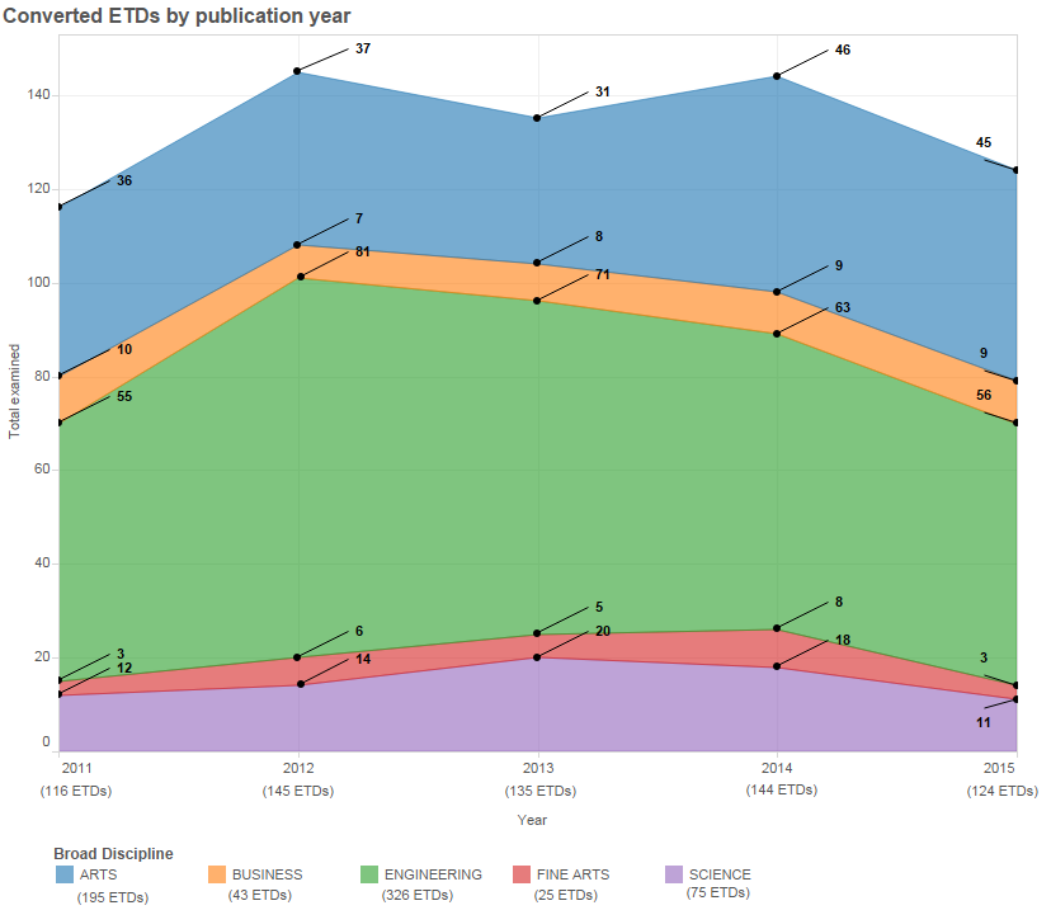


Figure 1. Converted ETDs by publication year and broad discipline

The 11,437 links extracted were checked for linkrot, each link accessed and its http response code recorded. 77% (8,834 of 11,437) of links returned an active 2xx http response code. 23% (2,603) of links could not be reached, returning a response code other than in the 2xx range. This includes 102 links in the 3xx range which failed to reach a destination after 50 redirects and were considered linkrot. Numbers of links, total link response, and link response by year broken down by broad discipline are shown in Figure 2, with accompanying data provided in Table 3 and discussed in the findings section.

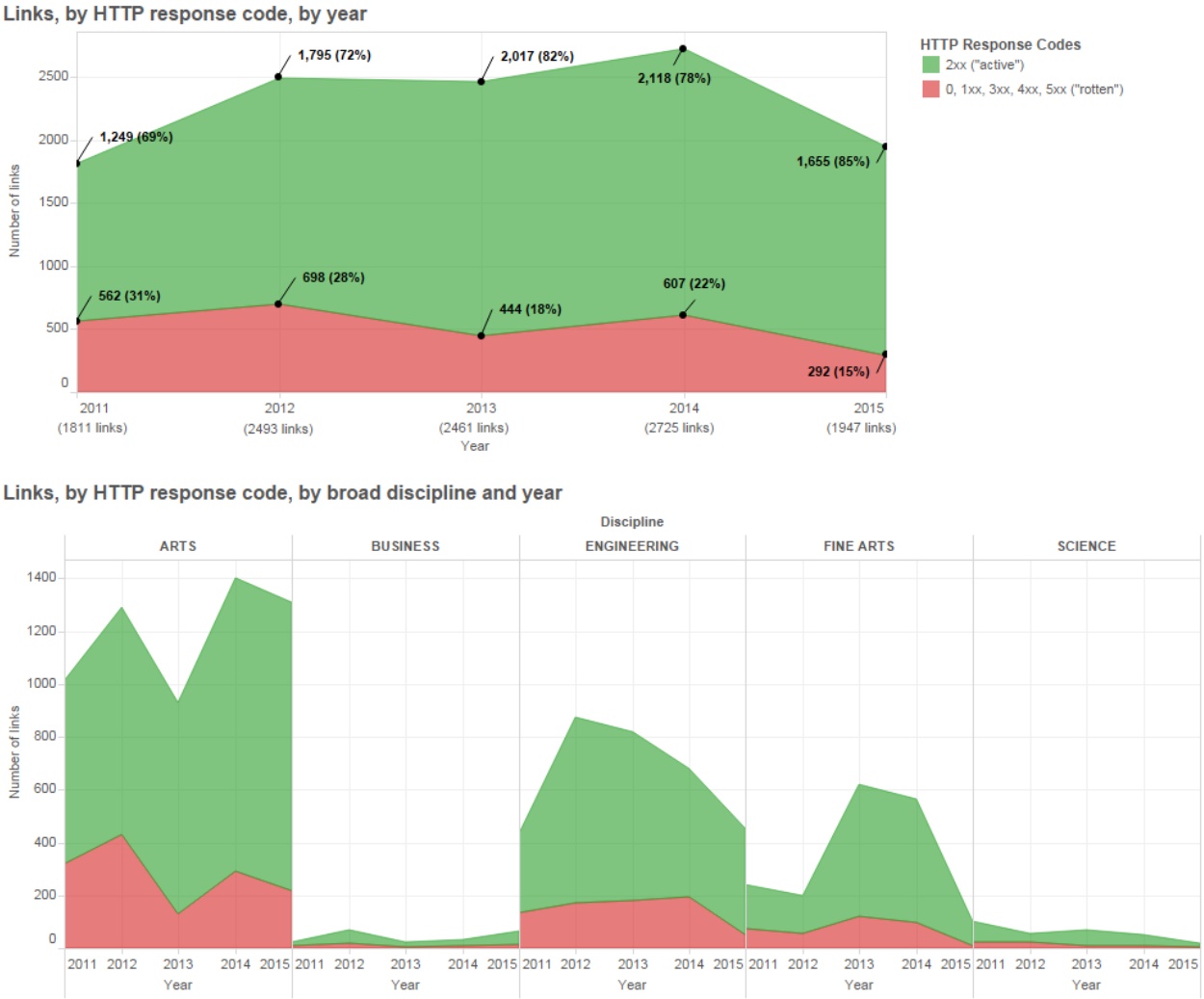


Figure 2. Link HTTP response codes, by broad discipline and year

Discipline	HTTP response code	2011	2012	2013	2014	2015	Total	% Active & Rotten**
Arts	2xx	691	864	800	1,108	1,093	4,556	77%
	all other*	320	428	131	293	218	1,390	23%
Business	2xx	14	52	17	22	50	155	74%
	all other	9	19	5	9	13	55	26%
Engineering	2xx	302	702	638	482	404	2,528	78%
	all other	134	172	180	196	49	731	22%
Fine Arts	2xx	165	143	504	467	94	1,373	79%
	all other	74	56	118	98	9	355	21%
Science	2xx	77	34	58	39	14	222	76%
	all other	25	23	10	11	3	72	24%
Subtotal	**2xx**	**1,249**	**1,795**	**2,017**	**2,118**	**1,655**	**8,834**	**77% active**
	all other	**562**	**698**	**444**	**607**	**292**	**2,603**	**23% rotten**
% Rotten		**31%**	**28%**	**18%**	**22%**	**15%**	**23%**	
Total		**1,811**	**2,493**	**2,461**	**2,725**	**1,947**	**11,437**	**100%**

Table 3. Breakdown by year and discipline showing active (2xx) and rotten (all others) response codes

*All other = 0, 1xx, 3xx (unresolved after 50 redirects), 4xx and 5xx response codes combined
** Active and rotten rates based on total links per discipline

METHODOLOGY — Content Drift

For the content drift phase, we wanted to sample documents from each of the five disciplines. ETDs which did not contain any links were excluded from the sample. Using only documents with one or more active links, a stratified random sample of 10% was drawn for a final sample of 49 ETDs containing a total of 990 links. A snippet of text surrounding each link was then also extracted from each ETD, along with any "date accessed" or "date viewed" information if present. Each link was manually visited, assessed for content drift, and observations recorded. The breakdown of the content drift sample is shown in Table 4.

Discipline	ETDs with links	ETDs with active links (2xx)	ETDs sampled for content drift*	Number of links extracted for sample
Arts	164	156	16	668
Business	31	28	3	12
Engineering	244	235	24	154
Fine Arts	23	23	2	136
Science	42	40	4	20
Total	**504**	**482**	**49**	**990**

Table 4. Breakdown of sample pool of ETDs for content drift analysis

* 10% sample drawn from each discipline's pool of ETDs; only ETDs with URLs relevant for content drift assessment.

Visited links were benchmarked against the existence of a memento -- an archived snapshot of that page located in the Wayback Machine.[37] Since the University sets a strict thesis submission deadline of 3 months prior to convocation, mementos prior to submission deadline would be sought. Based on the occurrences of "date accessed" and discursive information found in the snippets, we arrived at the supposition that links were likely to have been checked the closer the student approached final stages of manuscript preparation, although this is not verifiable. We set ourselves a soft window for locating an archived snapshot using a date 6 months prior to the convocation date as the benchmark; that is, for each semester's deadline date, an additional 3 months was added, arriving at a 6-months-prior-to-publication marker.

Since programmatic analysis of 990 links required time, expertise, and resources not available to us, we approached the problem heuristically. Assuming that online consultations are not linear, active links occurring multiple times in a document were given equal weight. Each link was manually checked in the Wayback Machine using "date viewed" if provided; if no date was provided (the majority of cases), Wayback was checked to see if an archived version existed as close to our 6 month soft marker as possible. If a memento was not found within a month earlier/later than the soft marker, then the nearest neighboring older memento was selected, if one existed. The original URL, the date the URL was visited, and whether a snapshot was located in Wayback was recorded. All links were checked during July-August 2016. If the initial web browser failed to access, a second and sometimes third browser was tried, using Safari, Chrome, and Internet Explorer (IE) in that order. Unsuccessful attempts to reach Wayback were rechecked in September. The question as to whether, and to what degree content drift had occurred was assessed, and is discussed in the next section.

FINDINGS AND DISCUSSION

Linkrot Findings

Of 664 ETDs examined for linkrot, 77% of links tested returned an active http response code in the 2xx range -- roughly three-quarters overall. Numbers of links by broad discipline varied greatly, as shown in Figure 2 (healthy links in green, linkrot shown in red). Linkrot rates ranged from 21% in Fine Arts, to 26% in Business, as seen in last column of Table 3. It should be noted that 2xx response codes are also returned for pages that disguise themselves as active links. For example, a URL returns an active status code when a domain has been parked (e.g. purchased to reserve the space), or when a customized 404-page-not-found is encountered. Since we had no mechanism in place to treat false positives, these were flagged during the linkrot phase as candidates for subsequent content drift analysis. 23% (2,604 of 11,437) of all links, returned a response code of something other than in the 2xx-range and considered linkrot -- roughly one-quarter. Response codes in the 4xx range alone, including 404-page-not-found errors, comprised 17% (1,916 of 11,437) of all links. Table 5 shows the breakdown of the total number of links that were visited in the spring of 2016 for linkrot determination.

HTTP response code category	Meaning of http response code*	Number of links	Percent of total links (%)
0	Empty response**	507	4%
1xx	Informational	2	0%
2xx	Successful	8,834	77%
3xx	Redirection†	102	1%
4xx	Client error	1,916	17%
5xx	Server error	76	1%
Total		**11,437**	**100%**

Table 5. Breakdown of HTTP response codes received

* We used http protocol definitions at http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html
** unofficial http response code due to request timing out
† failure to resolve after 50 redirects

Http responses ranged from a high of 85% active in 2015, to a low of 69% active in 2011, the oldest publication year. To put it differently, the most recent year exhibited a linkrot rate of 15%. Consistent with other studies, linkrot manifests itself quickly after publication and increases over time, as indicated by percentages shown in Figure 2.

Content Drift Findings

Of the 990 links visited to check for the presence of content drift, 764 (400 + 364), or 77%, had a Wayback memento compared 226 (92+134), or 23%, which did not. Slightly more than half of links with mementos, 52% (400 of 764), demonstrated some level of content drift when the

memento was compared to the current active link, while 48% (364 of 764) with mementos did not exhibit content drift. The presence of content drift by discipline, with/without mementos showing numbers of links tested, appears in Table 6.

Discipline	Number of links tested	**Content Drift detected**			**No Content Drift**		
		Memento found	Memento not found	Total	Memento found	Memento not found	Total
Arts	668	261	60	321	254	93	347
Business	12	5	0	5	4	3	7
Engineering	154	74	10	84	55	15	70
Fine Arts	136	55	22	77	38	21	59
Science	20	5	0	5	13	2	15
Total	**990**	**400**	**92**	**492**	**364**	**134**	**498**

Table 6. Presence of Content Drift by Discipline, with/without mementos

For links that had no memento in Wayback, content drift assessment was based on the presence of an observable date in the current active link, including copyright, and/or other details which positively correlated against our extracted snippet information. For example, some links retrieved a .pdf or other static file which correlated with the snippet, there being no reason to conclude its content had undergone change since publication, despite the lack of a memento. Snippets were also used in cases where a robots.txt file at the target URL had prevented Wayback from creating a memento. Occasional examination of the dissertation text was conducted to validate information extracted in the snippet. The 23% (226) which lacked mementos remain at significant risk and will fall prey to further drift as time passes.

As seen in Table 7, of 492 URLs manifesting content drift, 11% (54 of 492) were completely lost, linking to web domains that had been sold or were currently up for sale, and webpages replaced or removed. 9% (42 of 492) of web pages exhibited major change such that there was little correlation with snippets, or where website overhauls made assessment difficult, but not impossible. 36% (179 of 492) web links exhibited minor drift, primarily pages that differed somewhat from a memento in visual appearance, such as header and footer differences, changes in background theme or style, or changes in navigation or search functionality which did not represent a high degree of impairment. 7% (34 of 492) linked to continually updating websites, such as Wikipedia and news organizations, and 7% (35 of 492) were customized 404-page-not-found, distinctive enough to warrant separate categories. A full 30% (148 of 492) exhibited a multiplicity of changes of uncertain nature which we grouped together, such as pages where graphic or audio components had been removed or could not be retrieved, broken javascript that impeded access, browser failure, mementos not accessible after repeated attempts -- indicative of a range of issues affecting the quality of web archives and hence preservation.[38] The types of

content drift encountered, broken down by broad discipline and numbers of links, and percentage, is shown in Table 7.

Type of content drift	Arts	Business	Engineering	Fine Arts	Science	Total	% of type
Lost	45	0	3	6	0	54	11%
Major but findable	22	0	9	9	2	42	9%
Minor – redesigned but recognizable	128	2	30	17	2	179	36%
Ongoing updating website	25	3	5	0	1	34	7%
Custom 404	23	0	4	8	0	35	7%
Other	78	0	33	37	0	148	30%
Total	**321**	**5**	**84**	**77**	**5**	**492**	**100%**

Table 7. Types of content drift encountered, number of links by broad discipline

Though difficulties encountered during content drift assessment made further extrapolation problematic, the presence of reference rot was confirmed. Our 10% stratified random sample examined 990 active links, finding that roughly half (492 of 990) manifested some degree of content drift. For 364 links, or 36% overall, a benchmark memento was found and no content drift detected. Although many content drift changes can arguably be characterized as minor, it is not possible to ascertain where the content drift scale tips irremediably for any particular reader. What can be said with certainty is that 11% of active links which did not exhibit linkrot, and were quite live and accessible, fell into a small but unsettling group where the context of the cited web source is irrevocably lost. Of the 498 links which did not exhibit any evidence of content drift, 134, approximately one-third, have no memento archived and continue to remain at high risk.

A focused and deeper analysis of active links which might lead to a typology of content drift types would be a possible area of future study, though even the well-resourced study by Jones et al. which utilized a strict "ground truth" for comparing textual mementos over time, points out that classifying links would certainly be challenging.[39] A larger sample size might also allow closer analysis of disciplinary differences, which may lead to a better understanding of these types of content drift variations.

CONCLUSION

Reference rot in the form of linkrot and content drift were observed in ETDs in Spectrum, our institutional repository, and this confirmation should give pause for those charged with

stewardship of ETD collections. Theses and dissertations have long been viewed as material which contribute overall to academic scholarly output, and carry unique status within the academy. In August 2016, OpenDOAR registered 1600 institutional repositories with ETDs,[40] up from 1,100 institutions as reported in 2012 by grey literature specialist Schoepfel.[41]

Academic libraries have, in large part, facilitated the transition from paper to ETD with widespread adoption of institutional repository deposit programs, and along with that adoption comes a range of long-term preservation issues. Yet as Ohio State's Strategic Digital Initiatives Working Group pointed out, "Even in digital library communities, preservation all too often stands in for or is used interchangeably with byte level backup of content."[42] For long-term access, focus can productively be shifted to offset the immediate threat of incompleteness and inadequate capture.[43] Not much has changed since Hedstrom wrote back in 1997:

> *"With few exceptions, digital library research has focussed on architectures and systems for information organization and retrieval, presentation and visualization, and administration of intellectual property rights … The critical role of digital libraries and archives in ensuring the future accessibility of information with enduring value has taken a back seat to enhancing access to current and actively used materials."*[44]

Our understanding and discussion of digital preservation must be broadened, and attention turned to this key area of responsibility in the preservation life-cycle. The authors maintain that ETD content and link preservation is an editorial, not individual, imperative. Encouraging individual authors to perform their own archiving is doomed to fall short of even reasonable expectations. Instituting measures such as Perma, a distributed, redundant method of capturing and archiving web site content as part of the citation process must be pro-actively sought and built into library, and hence repository, workflows.[45] Browser plugins and automated solutions which use the Memento protocol for capturing and archiving web site content as part of the citation process do exist,[46] but naturally have to be implemented before they can take effect. Either way, efforts to operationalize existing mechanisms which are designed to reduce future loss would be extremely productive.

Responsibility for insuring not only current, but continuing future access to ETD content rests with those who maintain curatorial function of the repository. Academic librarians have assumed a prominent and *de facto* role as curators, facilitating the role of university publication and emphasizing its break away from previous ties with commercial entities. We collectively bear greater responsibility for this body of scholarly work, and need to move forward from a position of benign neglect to one of informed curation and pro-active preservation of an important collection of scholarly output which is at risk.

REFERENCES

1. Thomas H. Teper and Beth Kraemer, "Long-Term Retention of Electronic Theses and Dissertations," *College & Research Libraries* 63, no. 1 (January 1, 2002), 64, https://doi.org/10.5860/crl.63.1.61.

2 The term "reference rot" was introduced by the Hiberlink team. "Hiberlink – About," accessed March 31, 2016, http://hiberlink.org/about.html.

3. LOCKSS: Lots of Copies Keep Stuff Safe, accessed December 6, 2016, http://www.lockss.org/about/what-is-lockss/.

4. Mark Edward Phillips, Daniel Gelaw Alemneh, and Brenda Reyes Ayala, "Analysis of URL References in ETDs: A Case Study at the University of North Texas," *Library Management* 35, no. 4/5 (June 3, 2014), 294, https://doi.org/10.1108/LM-08-2013-0073.

5. Wallace Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," *Journal of the American Society for Information Science* 50, no. 2 (January 1, 1999): 162–80, https://doi.org/10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASI7>3.0.CO;2-B.

6. Wallace Koehler, "Web Page Change and Persistence—a Four-Year Longitudinal Study," *Journal of the American Society for Information Science & Technology* 53, no. 2 (January 15, 2002): 162–71, http://doi.org/10.1002/asi.10018.

7. Wallace Koehler, "A longitudinal study of Web pages continued: a consideration of document persistence." *Information Research* 9, no. 2 (2004): 9-2, http://www.informationr.net/ir/9-2/paper174.html.

8. Fatih Oguz and Wallace Koehler, "URL Decay at Year 20: A Research Note," *Journal of the Association for Information Science and Technology* 67, no. 2 (February 1, 2016): 477–79, https://doi.org/10.1002/asi.23561.

9. Mary F. Casserly and James Bird, "Web Citation Availability: Analysis and Implications for Scholarship," *College and Research Libraries* 64, no. 4 (July 2003): 300–317, http://crl.acrl.org/content/64/4/300.full.pdf.

10. Diomidis Spinellis, "The Decay and Failures of Web References," *Communications of the ACM* 46, no. 1 (January 2003): 71–77, https://doi.org/10.1145/602421.602422.

11. Carmine Sellitto, "A Study of Missing Web-Cites in Scholarly Articles: Towards an Evaluation Framework," *Journal of Information Science* 30, no. 6 (December 1, 2004): 484–95, https://doi.org/10.1177/0165551504047822.

12. Matthew E. Falagas, Efthymia A. Karveli, and Vassiliki I. Tritsaroli, "The Risk of Using the Internet as Reference Resource: A Comparative Study," *International Journal of Medical Informatics* 77, no. 4 (April 2008): 280–86, https://doi.org/10.1016/j.ijmedinf.2007.07.001.

13. Cassie Wagner et al., "Disappearing Act: Decay of Uniform Resource Locators in Health Care Management Journals," *Journal of the Medical Library Association* 97, no. 2 (April 2009): 122–30, https://doi.org/10.3163/1536-5050.97.2.009.

14. Robert Sanderson, Mark Phillips, and Herbert Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento," *arXiv:1105.3459 [Cs]*, May 17, 2011, http://arxiv.org/abs/1105.3459.

15. Jonathan Zittrain, Kendra Albert, and Lawrence Lessig, "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations," *Legal Information Management* 14, no. 2 (June 2014): 88–99, https://doi.org/10.1017/S1472669614000255.

16. "Hiberlink - About," accessed March 31, 2016, http://hiberlink.org/about.html.

17. "Hiberlink - Our Research," accessed March 31, 2016, http://hiberlink.org/research.html.

18. Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin. "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot," *PLoS One* 9, no. 12 (December 26, 2014), https://doi.org/10.1371/journal.pone.0115253.

19. Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, Claire Grover. "Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content," *PLOS ONE* 11, no. 12 (December 2, 2016): e0167475, https://doi.org/10.1371/journal.pone.0167475.

20. Martin Halbert, Katherine Skinner, and Matt Schultz, "Preserving Electronic Theses and Dissertations: Findings of the Lifecycle Management for ETDs Project," Text, (August 6, 2015), 2, http://educopia.org/presentations/preserving-electronic-theses-and-dissertations-findings-lifecycle-management-etds.

21. For a recent overview, see Sarah Potvin and Santi Thompson, "An Analysis of Evolving Metadata Influences, Standards, and Practices in Electronic Theses and Dissertations," *Library Resources & Technical Services* 60, no. 2 (March 31, 2016): 99–114, https://doi.org/10.5860/lrts.60n2.99.

22. Joy M. Perrin, Heidi M. Winkler, and Le Yang, "Digital Preservation Challenges with an ETD Collection — A Case Study at Texas Tech University," *The Journal of Academic Librarianship* 41, no. 1 (January 2015): 98–104, https://doi.org/10.1016/j.acalib.2014.11.002.

23. Sanderson, Phillips, and Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento," http://arxiv.org/abs/1105.3459.

24. Phillips, Alemneh, and Ayala, "Analysis of URL references," https://doi.org/10.1108/LM-08-2013-0073.

25. Alfred S. Sife and Ronald Bernard, “Persistence and Decay of Web Citations Used in Theses and Dissertations Available at the Sokoine National Agricultural Library, Tanzania,” *International Journal of Education and Development Using Information and Communication Technology* 9, no. 2 (2013): 85–94, http://eric.ed.gov/?id=EJ1071354.

26. “ETD2014 — University of Leicester,” *University of Leicester*, accessed January 27, 2016, http://www2.le.ac.uk/library/downloads/etd2014.

27. EDINA, University of Edinburgh, “Reference Rot: Threat and Remedy,” (Education, 04:54:38 UTC), http://www.slideshare.net/edinadocumentationofficer/reference-rot-and-linked-data-threat-and-remedy.

28. Peter Burnhill, Muriel Mewissen, and Richard Wincewicz, “Reference Rot in Scholarly Statement: Threat and Remedy,” *Insights the UKSG Journal* 28, no. 2 (July 7, 2015): 55–61, https://doi.org/10.1629/uksg.237.

29. Concordia University University Graduate Programs, accessed April 7, 2016, http://www.concordia.ca/academics/graduate.html.

30. Klein et al., "Scholarly Context Not Found," https://doi.org/10.1371/journal.pone.0115253.

31. Ke Zhou, Richard Tobin, and Claire Grover, “Extraction and Analysis of Referenced Web Links in Large-Scale Scholarly Articles,” in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’14 (Piscataway, NJ, USA: IEEE Press, 2014), 451–452, http://dl.acm.org/citation.cfm?id=2740769.2740863.

32. Pdftohtml v0.38 win32, meshko (Mikhail Kruk), http://pdftohtml.sourceforge.net/ accessed September 20, 2015. (Actual download is at http://sourceforge.net/projects/pdftohtml/).

33. Give me text! Open Knowledge International, accessed October 26, 2015-March 7, 2016, http://givemetext.okfnlabs.org/.

34. Phillips, Alemneh, and Ayala, "Analysis of URL references," https://doi.org/10.1108/LM-08-2013-0073.

35. “In Search of the Perfect URL Validation Regex,” accessed December 7, 2015, https://mathiasbynens.be/demo/url-regex. We selected "@gruber v2" for our extraction.

36. cURL v7.45.0, "command line tool and library for transferring data with URLs," accessed October 18, 2015, http://curl.haxx.se/.

37. We have used the term "memento" in lowercase to denote a snapshot souvenir page, to distinguish from an automated service utilizing the Memento protocol.

38. For a good overview of the types of problems, see Michael L. Nelson, Scott G. Ainsworth, Justin F. Brunelle, Mat Kelly, Hany SalahEldeen and Michele Weigle, "Assessing the Quality of Web Archives" 1 vol., *Computer Science Presentations*, Book 8 (Old Dominion University. ODU Digital Commons, 2014). http://digitalcommons.odu.edu/computerscience_presentations/8.

39. Shawn M. Jones, et al. "Scholarly Context Adrift," https://doi.org/10.1371/journal.pone.0167475.

40. OpenDOAR search of Institutional Repositories with Theses at http://www.opendoar.org/find.php, accessed August 26, 2016.

41. Joachim Schöpfel, "Adding value to electronic theses and dissertations in institutional repositories." *D-Lib Magazine* 19, no. 3 (2013): 1. https://doi.org/10.1045/march2013-schopfel.

42. Strategic Digital Initiatives Working Group. *Implementation of a Modern Digital Library at The Ohio State University*. (Apr 2014). https://library.osu.edu/documents/SDIWG/sdiwg_white_paper.pdf. (Published).

43. Tim Gollins. "Parsimonious Preservation: Preventing Pointless Processes! (The Small Simple Steps That Take Digital Preservation a Long Way Forward)," in *Online Information Proceedings* UK National Archives, 2009. Available at http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf.

44. Margaret Hedstrom, "Digital preservation: a time bomb for digital libraries." *Computers and the Humanities* 31, no. 3 (1997): 189-202. https://doi.org/10.1023/A:1000676723815.

45. Zittrain, Albert, and Lessig, "Perma," https://doi.org/10.1017/S1472669614000255.

46. Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila L. Balakireva, Scott Ainsworth, and Harihar Shankar, "Memento: Time Travel for the Web," *arXiv:0911.1112 [Cs]*, November 5, 2009, http://arxiv.org/abs/0911.1112.