ARCHIVABILITY OF DOCUMENTS IS EMERGING AS A BEST PRACTICE IN DIGITAL PRESERVATION. WHAT'S IN YOUR INSTITUTIONAL REPOSITORY?

REFERENCE ROT : a digital preservation issue beyond file formats

Mandated electronic deposit of theses and dissertations (ETDs) carries with it digital preservation concerns for librarians in a new role as defacto digital publisher.
As scholarly content vanishes due to the nature of the ephemeral web, what's next for digital-born documents deposited in our institutional repositories?

LINK ROT + CONTENT DRIFT ∴ MEMENTOS

LINK ROT

Links pointing to webpages & resources are no longer available at URL address, e.g., 404-page not found.

"All three corpora show a moderate, yet alarming, **link rot ratio for references made in recent articles, published in 2012: 13% for arXiv, 22% for Elsevier, and 14% for PMC** ... Going back to the **earliest publication year in our corpora, 1997, the ratios become 34%, 66%, and 80%**, respectively."

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), e115253.

CONTENT DRIFT

Links work, but URL page content has evolved over time and differs, sometimes dramatically, from what was there originally.

"We find that **for over 75% of references the content has drifted away** from what it was when referenced."

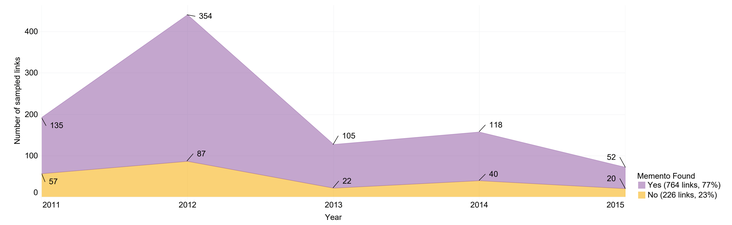
Jones, S. M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & Grover, C. (2016). Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PloS one, 11*(12), e0167475.

MEMENTOS

Digital snapshots, i.e., screen captures, which are preserved in publicly accessible archives.

The international archiving community has been periodically crawling websites and saving mementos for years. An example of such **incidental archiving** is Internet Archive's Wayback Machine. Even if libraries have undertaken no actions to insure digital preservation, some mementos will exist, as our research shows.

Sampled Links (990), by Memento Found, by Year

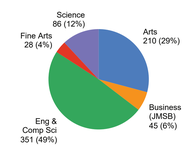


CONCORDIA UNIVERSITY'S SPECTRUM RESEARCH REPOSITORY

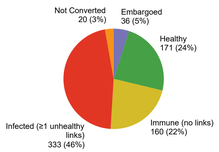
To see if Concordia University's ETDs suffered from reference rot, we examined PhD dissertations deposited in Spectrum during a 5 year period (Spring 2011 - Fall 2015)
- Documents were downloaded, converted to text and mined for URLs
- URLs were checked using cURL to obtain http response codes

Using a 10% stratified random sample of 990 links, we found:
- About *half* of PhD links sampled (492/990) exhibited content drift
- 77% (764/990) had mementos, 23% (226/990) had no memento
- Content for 11% of sampled links (54/990) is *lost and not recoverable*

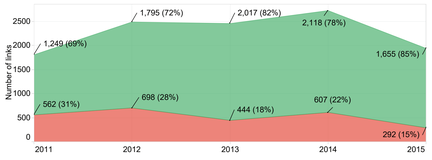
Total PhDs (720), by Discipline



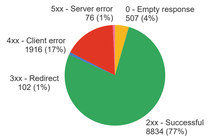
Total PhDs (720), by Document Health



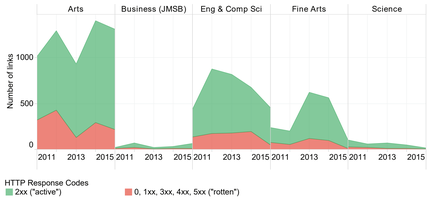
Links, by HTTP Response Code, by Year



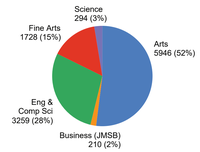
Link Distribution, by HTTP Status Code (11,437 links)



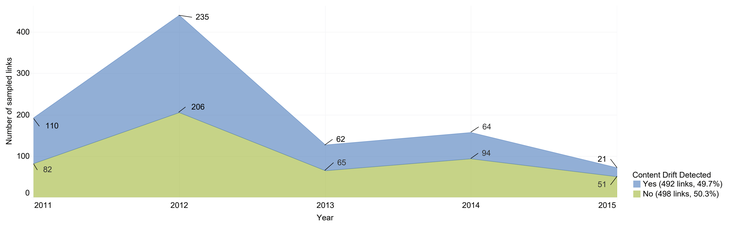
Links, by HTTP Response Code, by Discipline and Year



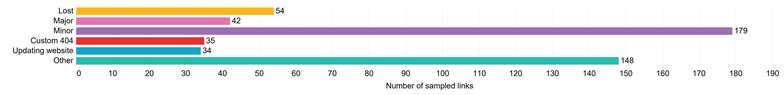
Total Links (11,437), by Discipline



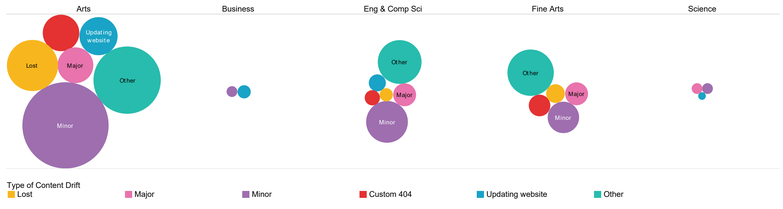
Sampled Links (990), by Content Drift Detected, by Year



Total Content Drift (492), by Type of Drift



Total Content Drift (492), by Type of Drift, by Discipline



WHAT ABOUT THE GAPS?

Reframing Our Responsibilities

- Avoid use of URL shorteners (e.g., bit.ly)
- Make repositories and publishing websites archive-friendly
- Add archiving crawlers to whitelists
- Collaborate with Thesis Office to systematically preserve ETD links

MAKE AND SAVE MEMENTOS

INTERNET ARCHIVE — INTERNET ARCHIVE WayBackMachine

- Use Save Page Now

- Install browser extension

Archive-it: Internet Archive's web archiving subscription service. Save collections; rescue websites before they disappear.

Get to know perma.cc

- Individual/Institutional accounts
- Install browser extension

Massicotte, M., & Botter, K. (2017). Reference rot in the repository: A case study of electronic theses and dissertations (ETDs) in an academic library. *Information Technology and Libraries (ITAL), 36*(1), 11-28.

UNIVERSITÉ Concordia UNIVERSITY LIBRARY

Mia.Massicotte@concordia.ca
Kathleen.Botter@concordia.ca