Reference rot analysis of PhD theses in Spectrum, Concordia's Institutional Repository

14th Annual Research Forum April 29, 2016

Kathleen Botter, Systems Librarian, Concordia University Library

Mia Massicotte, Systems Librarian, Concordia University Library

Long-term access to the scholarly record

One of our major challenges is preservation

- text must be stable
- citations and references must be retrievable
- a digital library "must remain a repository of fact in addition to being a dissemination point of information"

-- Dimitrova and Bugeja. Consider the source: predictors of online citation permanence in communication journals. *Portal: Libraries and the Academy*. 6(3), 2006. p.270

Why do we care?

- research builds upon previous work of others; verifiable evidence
- referencing is a manifestation of academic scholarship
- quality and thoroughness substantiated via attribution of sources
- current as well as future access to born digital is in jeopardy unless preservation is taking place

Why electronic theses and dissertations (ETDs)?

- electronic deposit is mandated by parent institution
- should remain stable over time
- library maintains stewardship of thesis collection
- scholarly content which is born-digital

Hard to pin down digital

- different kinds of digital content
- websites come and go
- page content is in flux

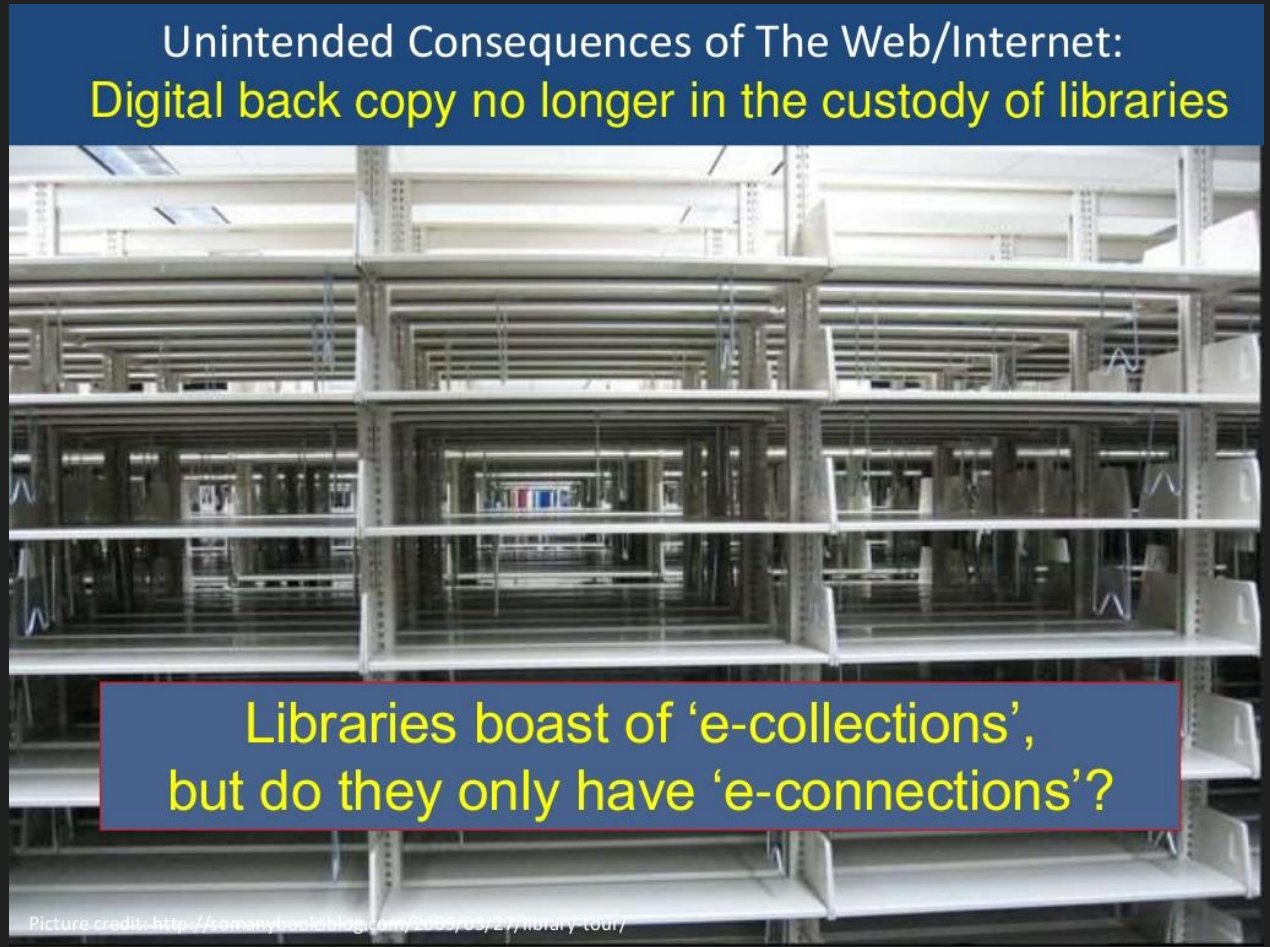
- digital scholarly content is susceptible to loss
- collections are in jeopardy unless preservation is taking place

Where is preservation taking place?

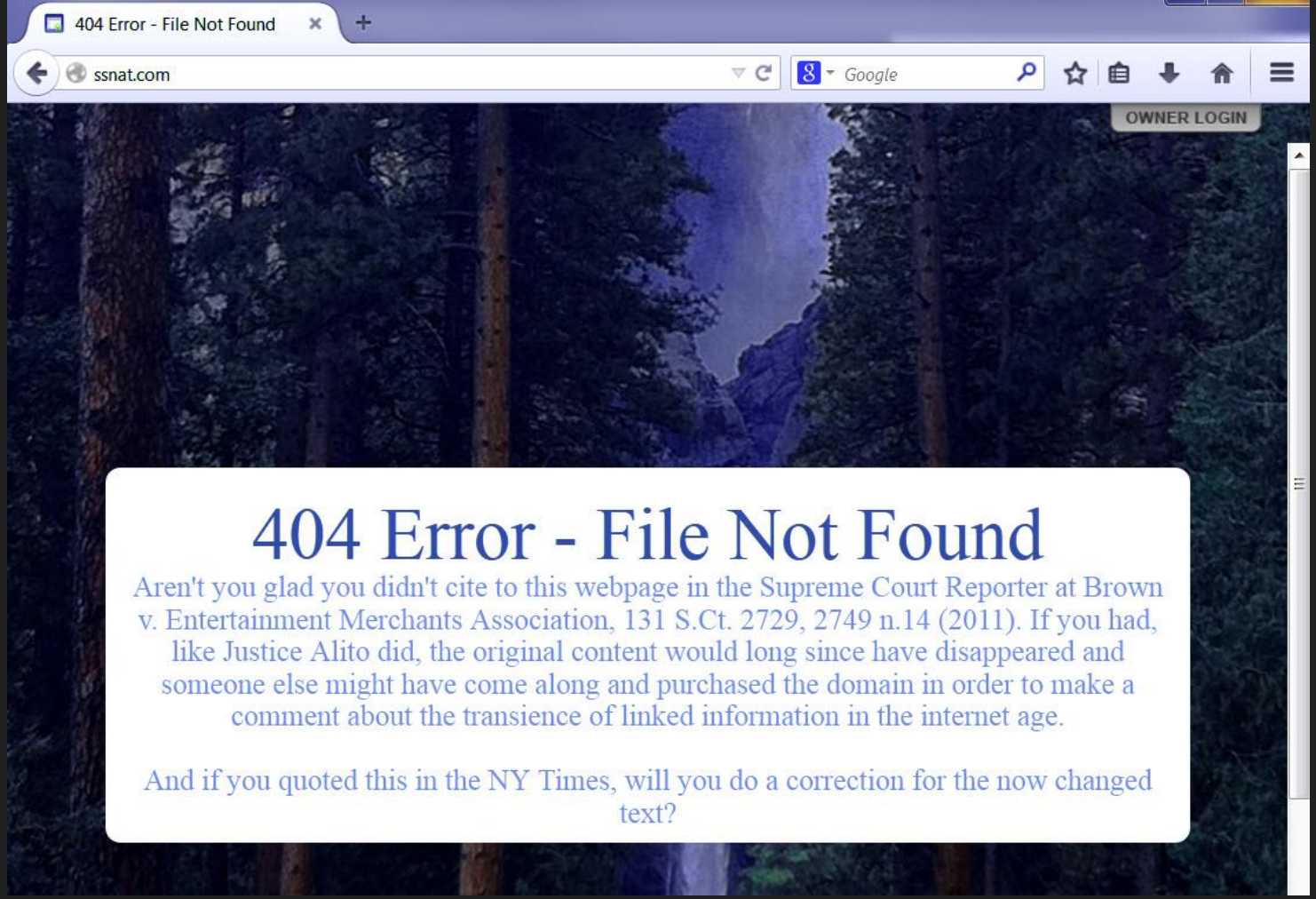
In his 2015 piece on *Preserving the born-digital record,* James Neal bluntly states:

WE ARE IN TROUBLE.

* Preserving the born-digital record: many more questions than answers. *American Libraries*, May 28, 2015 http://americanlibrariesmagazine.org/2015/05/28/preserving-the-born-digital-record/



Peter Burnhill, Standing on the digits of giants: research data, preservation and innovation. ALPSP/DPC, London, UK, 8 March 2016. http://image.slidesharecdn.com/alpsp2016hiberlinkburnhill-forpdf-160317145638/95/ensuring-the-integrity-continuity-of-our-record-of-scholarship-7-1024.jpg?cb=1458226778


404 Error - File Not Found
ssnat.com
Google
OWNER LOGIN
404 Error - File Not Found
Aren't you glad you didn't cite to this webpage in the Supreme Court Reporter at Brown v. Entertainment Merchants Association, 131 S.Ct. 2729, 2749 n.14 (2011). If you had, like Justice Alito did, the original content would long since have disappeared and someone else might have come along and purchased the domain in order to make a comment about the transience of linked information in the internet age.
And if you quoted this in the NY Times, will you do a correction for the now changed text?

Reference Rot and Law

Harvard team studied link permanence in 2012

- 3 Harvard law and policy-related publications (1996-2012)

  70% of links suffer reference rot

- database of 555 links U.S. Supreme Court opinions

  50% of links suffer reference rot*

*Zittrain J., Albert K., & Lessig L. (2014). Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Legal Information Management 14(2)*, 88-99. doi: 10.1017/S1472669614000255

Law library community takes action

Harvard study prompted the creation of

- Perma.cc (2013)
  - shared digital archiving service maintained and used by network of law libraries
- Hiberlink Project (2013-2015)
  - two-year funded project to study ‘reference rot’

What is Reference Rot?

Term coined in 2013 by Hiberlink Team

Two characteristics: Link Rot + Content Drift

- Link Rot = "404 Page Not Found"
- Content Drift = change in page content over time

Reference Rot in STM

Landmark study (2014) determined 1 in 5 STM articles suffers from reference rot

- science, technology and medical articles, 1997-2012
- 3.5 million articles collected
- 3 sources: arXiv, Elsevier, Pubmed Central
- 1.8 million articles with open web references analyzed:

  → of those, 7/10 suffer reference rot*

*Klein M., Van de Sompel H., Sanderson R., Shankar H., Balakireva L., Zhou K., et al. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE 9(12),* e115253. doi:10.1371/journal.pone.0115253

Research Questions

- What percentage of links in Concordia's ETD collection are affected by reference rot?
- Which disciplines exhibit more reference rot than others?
- What can be done to mitigate the problem?

Spectrum Research Repository

- Concordia University's institutional repository
  http://spectrum.concordia.ca
- scope limited to PhD dissertations
- started with first mandatory deposit year, Spring 2011
- 5 years, Spring 2011-Fall 2015, 719 total dissertations

What to look for

Using the Hiberlink typology, documents (dissertations) can be immune, healthy or infected:

- Immune: dissertation contains no links
- Healthy: contains links, all of which are active
- Infected: one or more links cannot be resolved

Methodology

Sequence for Extracting and Testing Links

1. download and convert pdfs to xml
2. use *regular expression* on each converted file to find links
3. manually verify (and fix) links
4. use cURL to get http status code for each link

   → output: original URL, final URL, http status code

Selecting a Regular Expression

How to identify a link?

- Search for known top level domains (TLD) e.g. .com, .edu)?
- Search for only well-formed web links e.g. start with http://?

→ Gruber v2 aka Daring Fireball

"intended to match any URLs, including 'mailto:foo@example.com', 'x-whatever://foo', etc."
(https://gist.github.com/gruber/249502)

Gruber v2 Regular Expression

(?i)\b((?:[a-z][\w-]+:(?:/{1,3}|[a-z0-9%])|www\d{0,3}[.]
|[a-z0-9.\-]+[.][az]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]
+|(\([^\s()<>]+\)))*\))
+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))*\)
|[^\s`!()\[\]{};:'".,<>?«»“”‘’]))

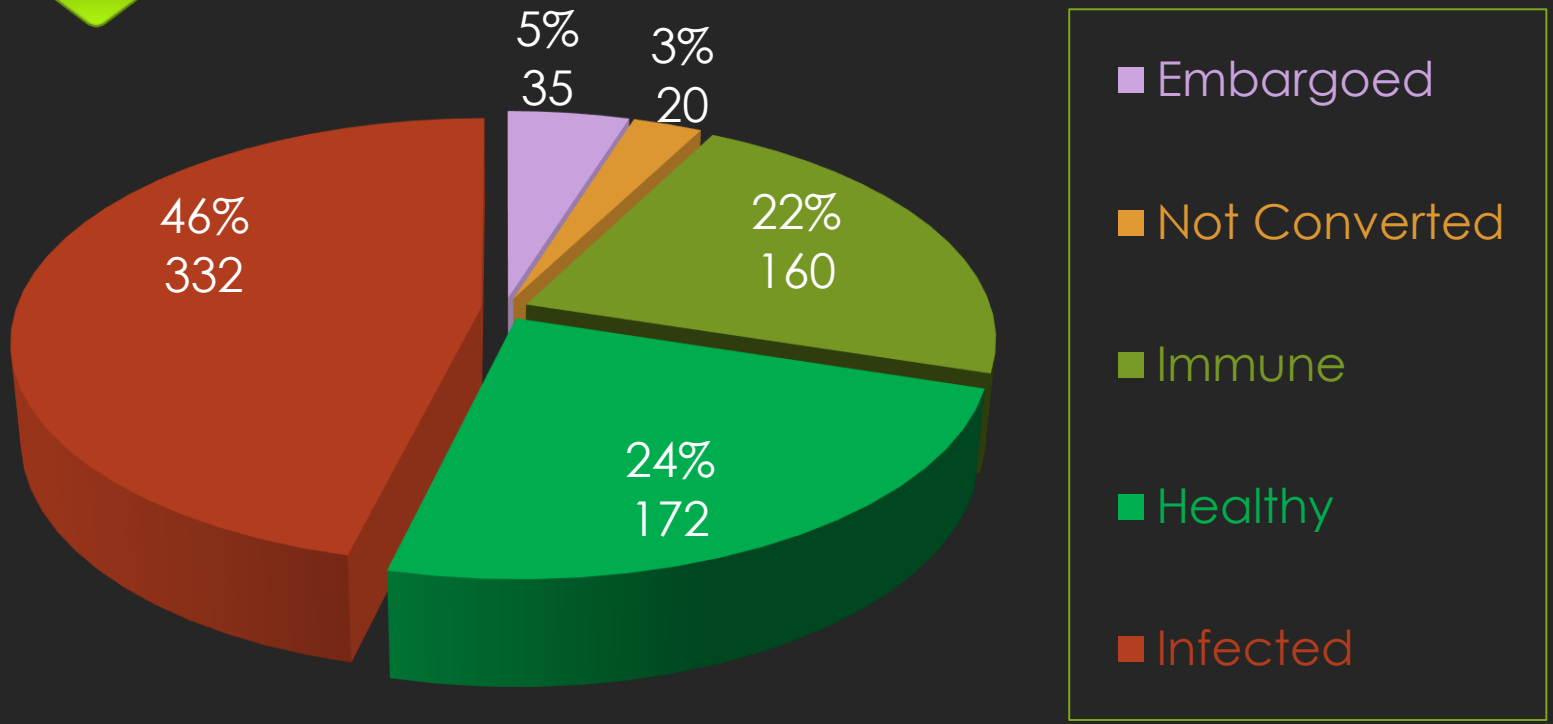
Challenges

- not all pdfs could be converted to xml
- inconsistent formatting of links in dissertations
  - ellipses (…) in long links
- typos, and less obvious mistakes
  - missing .edu, :, //

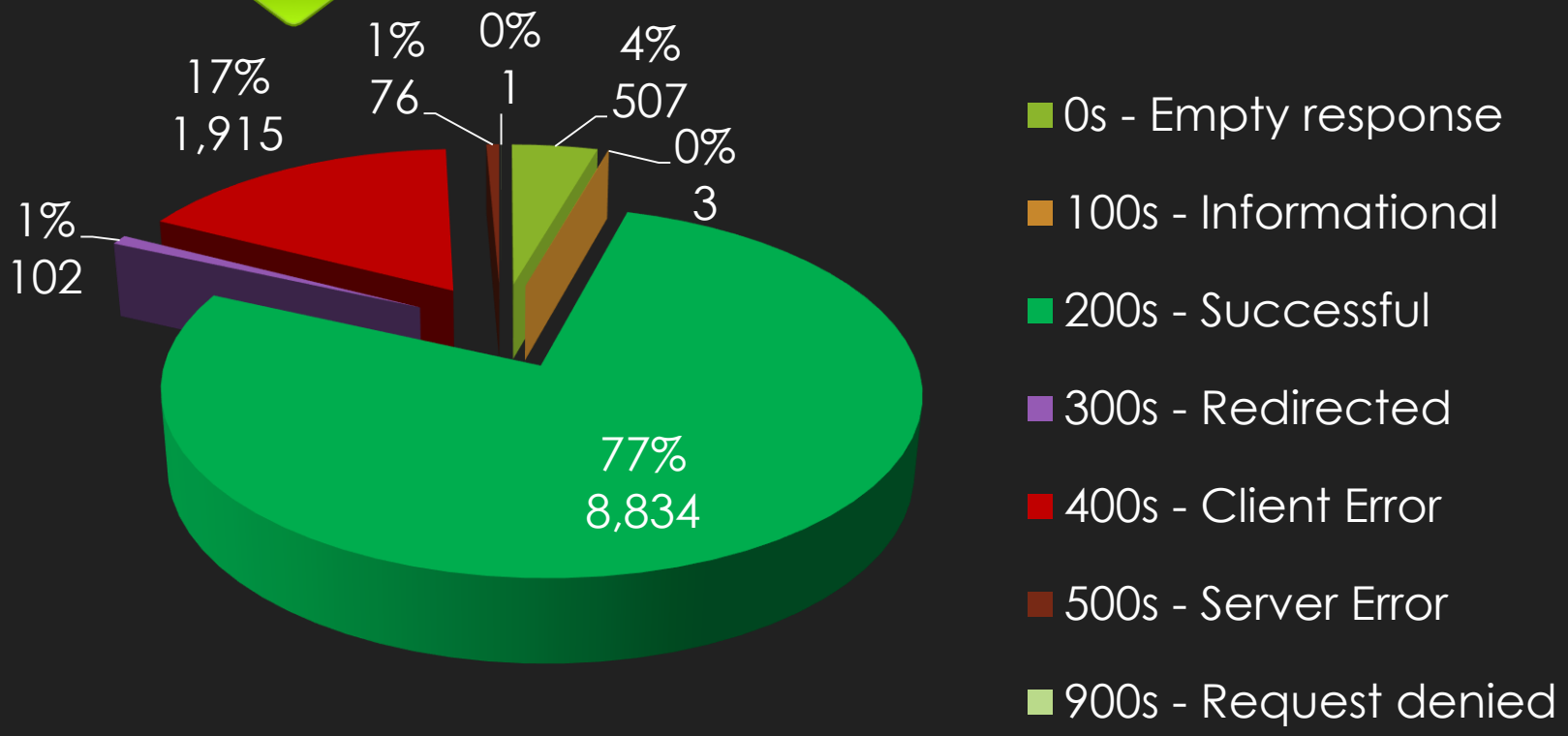
Preliminary Results

- 719 PhD dissertations
- 11,438 links

719 Total PhDs (2011-2015)

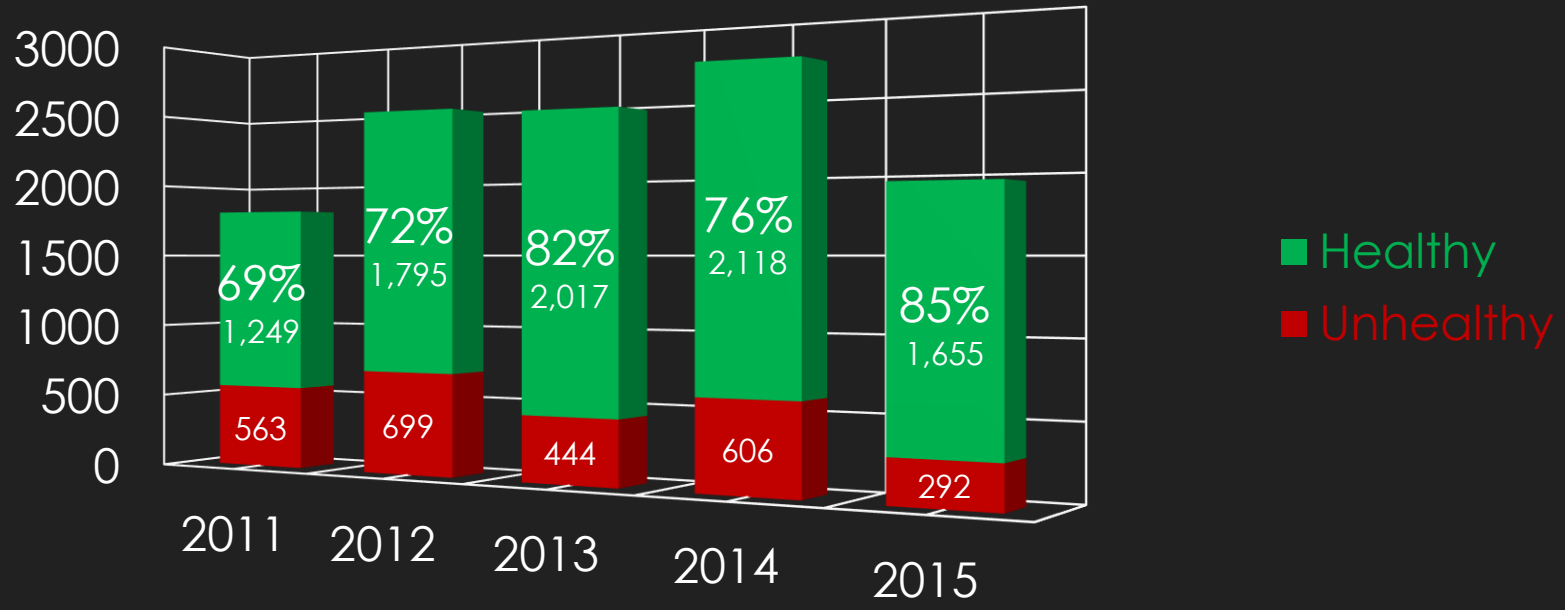


Link Distribution by HTTP status code (11,438 links)

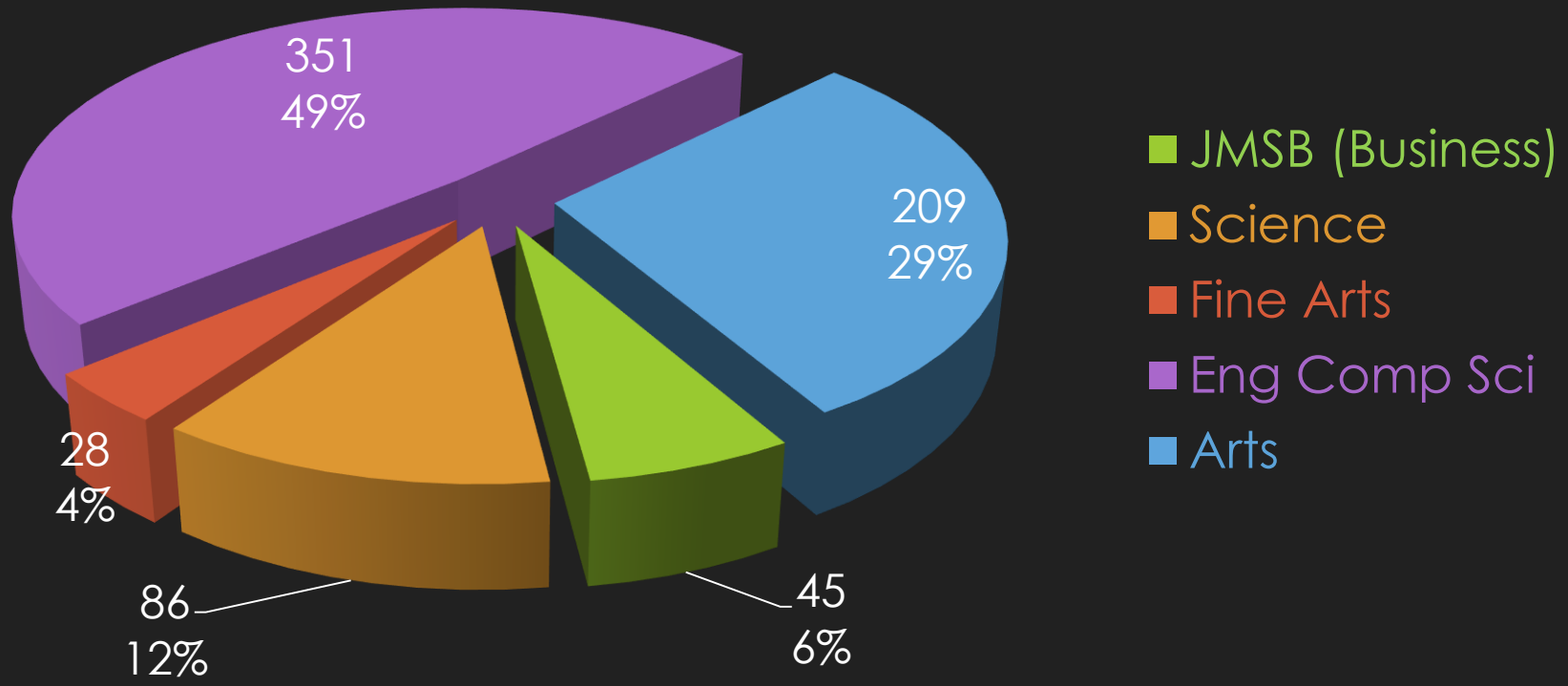


Healthy Links / By Year

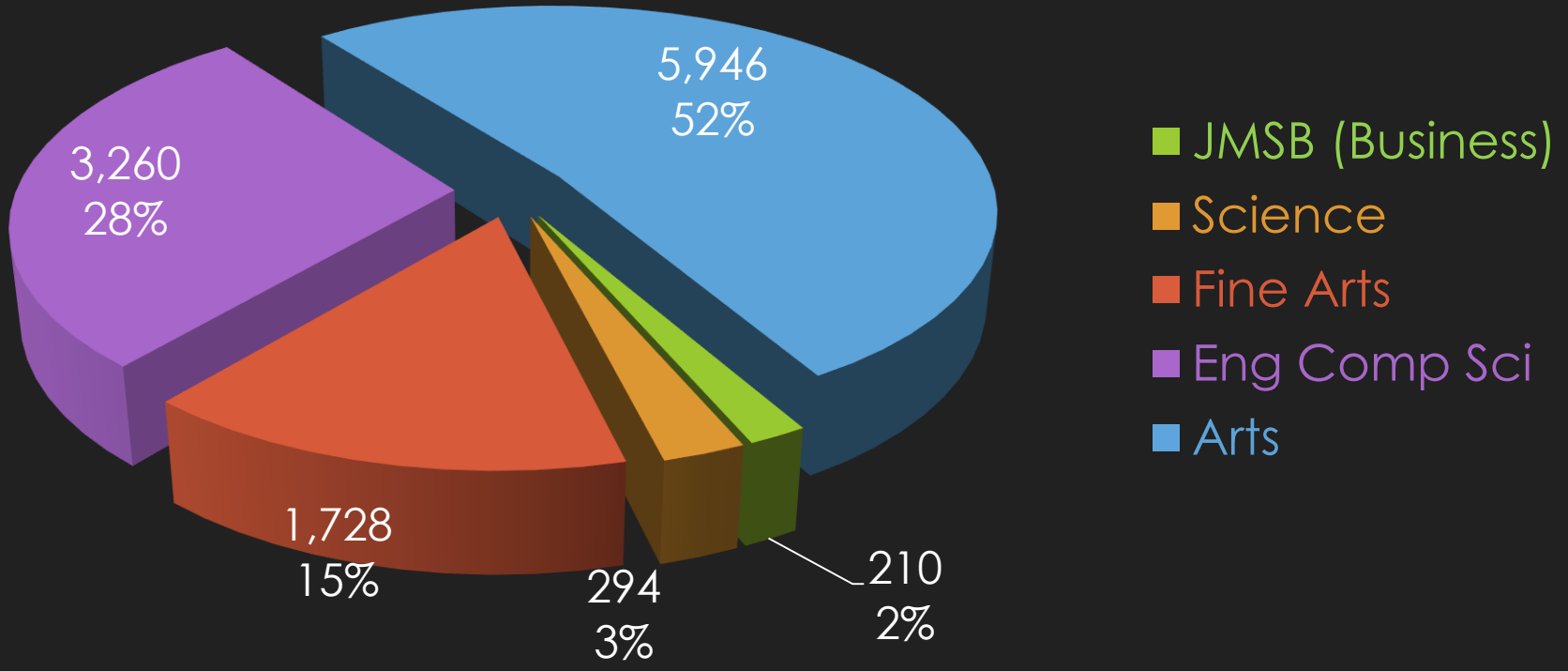
8,834 (77%) of 11,438 total links healthy



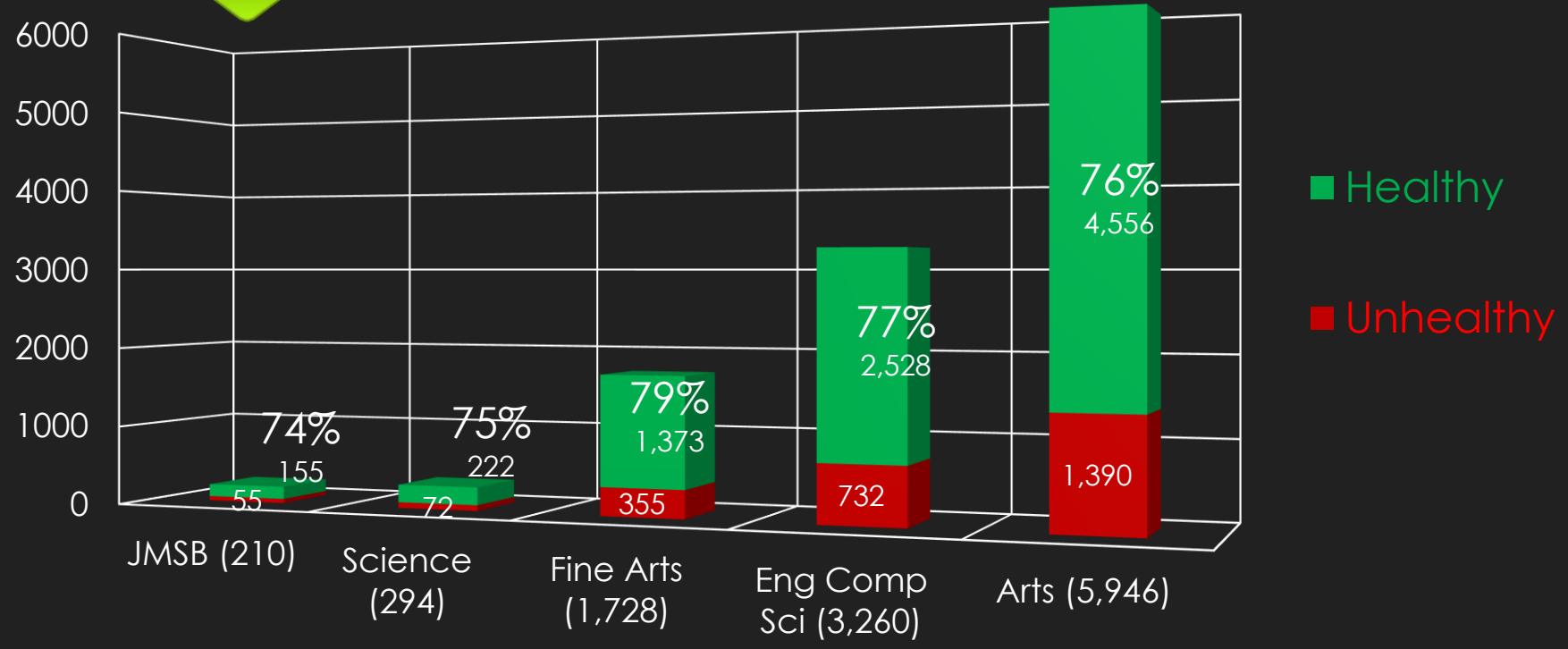
Total PhDs (719) by Discipline



Total links (11,438) by Discipline



Total Link Health / by Discipline



Surprise!

3,143 DOIs found

- 68% (2,146) DOIs appeared in last two years examined
- 42% (1,322) DOIs appeared in last year alone (2015)

Next steps: Content Drift

- in progress: sampling healthy links
- assessing Content Drift in healthy links
- visit link and compare to mementos

Early evidence of Content Drift

- evidence of “custom 404” in link rot results

original URL: http://www.gfkrt.com/imperia/md/content/rt-france/cp_gfk_ march___de_la_bd_-_39eme___dition_de_la_fibd.pdf

final URL: http://www.gfk.com/404/

http status code: 200

Mitigating Reference Rot

Mementos

- digital “snapshot in time” of a webpage
- examples include:
  - Internet Archive’s Wayback Machine
  - Archive-It
  - archive.today
- incidental archiving

Mitigation at the source

- browser plugins (e.g. Zotero) that can create mementos
- referencing mementos in the citation
  - proposed addition of 2 HTML elements:
    <a href="http://library.concordia.ca"
    data-versionurl=http://archive.today/IX5vo
    data-versiondate="2014-04-17">
- better editorial review of link formatting
  - avoid shortened URLs (bit.ly)

Institution / Library's Role

Should mementos be part of document deposit process?

- Who creates the memento?

- Who owns / guarantees the mementos existence?
  - copyright issues
  - public vs. institutional memento archives
  - national initiatives

References

1. Dimitrova,D. and M. Bugeja. Consider the source: predictors of online citation permanence in communication journals. *Portal: Libraries and the Academy*. 6(3), 2006. p.270
2. Neal,J. Preserving the born-digital record: many more questions than answers. *American Libraries*, May 28, 2015 http://americanlibrariesmagazine.org/2015/05/28/preserving-the-born-digital-record/
3. Burnhill, P. Standing on the digits of giants: research data, preservation and innovation. Presentation at ALPSP/DPC, London, UK, 8 March 2016. http://image.slidesharecdn.com/alpsp2016hiberlinkburnhill-forpdf-160317145638/95/ensuring-the-integrity-continuity-of-our-record-of-scholarship-7-1024.jpg?cb=1458226778
4. Zittrain J., Albert K., & Lessig L. (2014). Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Legal Information Management 14(2)*, 88-99. doi: 10.1017/S1472669614000255
5. Perma:cc http://perma.cc/
6. Klein M., Van de Sompel H., Sanderson R., Shankar H., Balakireva L., Zhou K., et al. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE 9(12),* e115253. doi:10.1371/journal.pone.0115253
7. Gruber, J. Liberal, Accurate Regex Pattern for Matching Web URLs. http://gist.github.com/gruber/249502
8. Chrome Memento Time Travel plugin for Zotero http://chrome.google.com/webstore/detail/memento-time-travel/jgbfpjledahoajcppakbgilmojkaghgm

Thank you

Questions?

Kathleen Botter, Systems Librarian
Concordia University Library
Kathleen.Botter@concordia.ca

Mia Massicotte, Systems Librarian
Concordia University Library
Mia.Massicotte@concordia.ca