Text & Data Mining Clauses in Academic Library Licenses: A Case Study

Paul Grewal and Kirsten Huhn

Concordia University Library

April 29, 2016

Outline

- Introduction
- Definition
- What Researchers are Saying
- Role of Academic Libraries in supporting TDM research
- Our study: Research Questions
- Our study: Methodology
- Our study: Preliminary Results
- Our study: Discussion/Conclusions
- List of References

Introduction

- The world is drowning in data
- Text and Data Mining (TDM) as a research technique offers a way to analyze large datasets quickly to offer new insights
- TDM is increasing in popularity across many disciplines, leading to new types of research

Introduction

- Many researchers use their own datasets or OA datasets
- Libraries pay for access to large datasets (including full-text, abstract, citation databases with their metadata)
- Access is often restricted for licensing, copyright, and technical reasons
- We want to look at the terms of our current licenses to see how we can better support our researchers in accessing paid-for data sets.

CRKN Model License Definition

A machine process by which information may be derived by identifying patterns and trends within natural language through text categorization, statistical pattern recognition, concept or sentiment extraction, and the association of natural language with indexing terms. (Canadian Research Knowledge Network, 2014)

What Researchers Are Saying

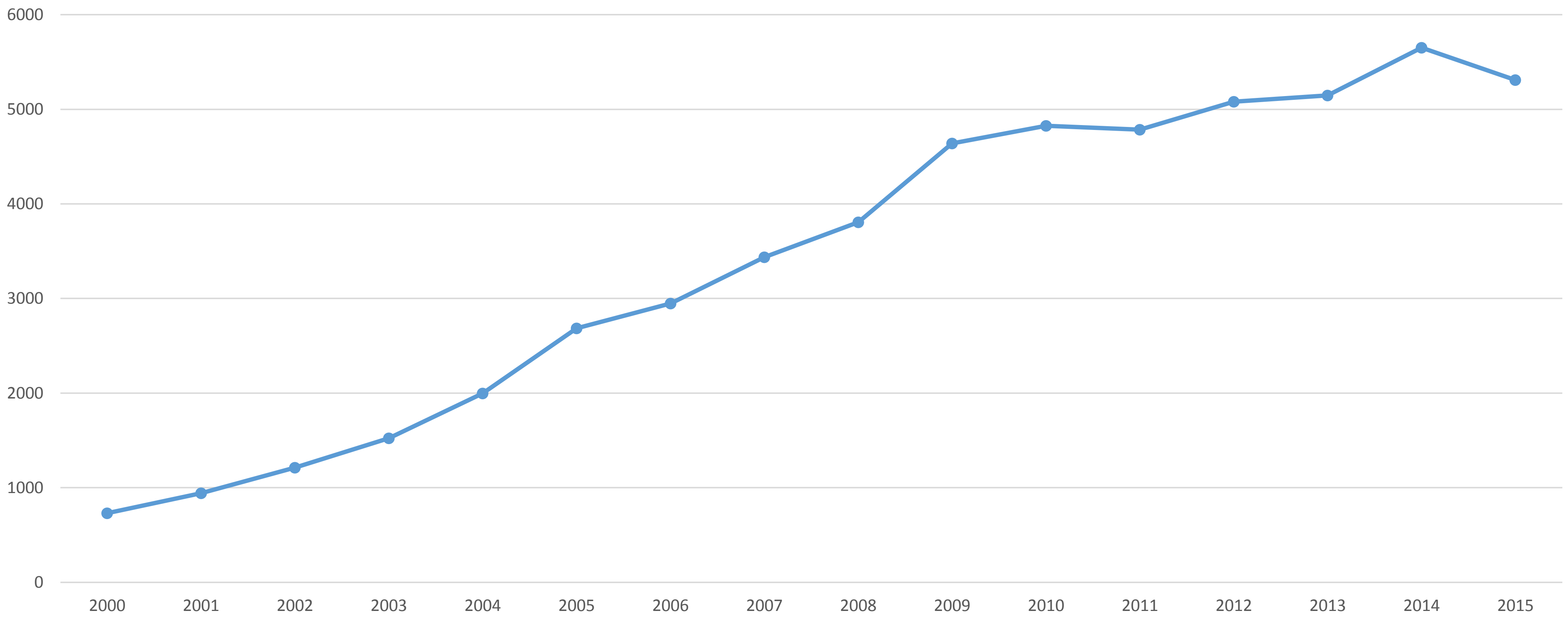
Text Mining the Novel Project (.txtLAB@McGill, n.d.)

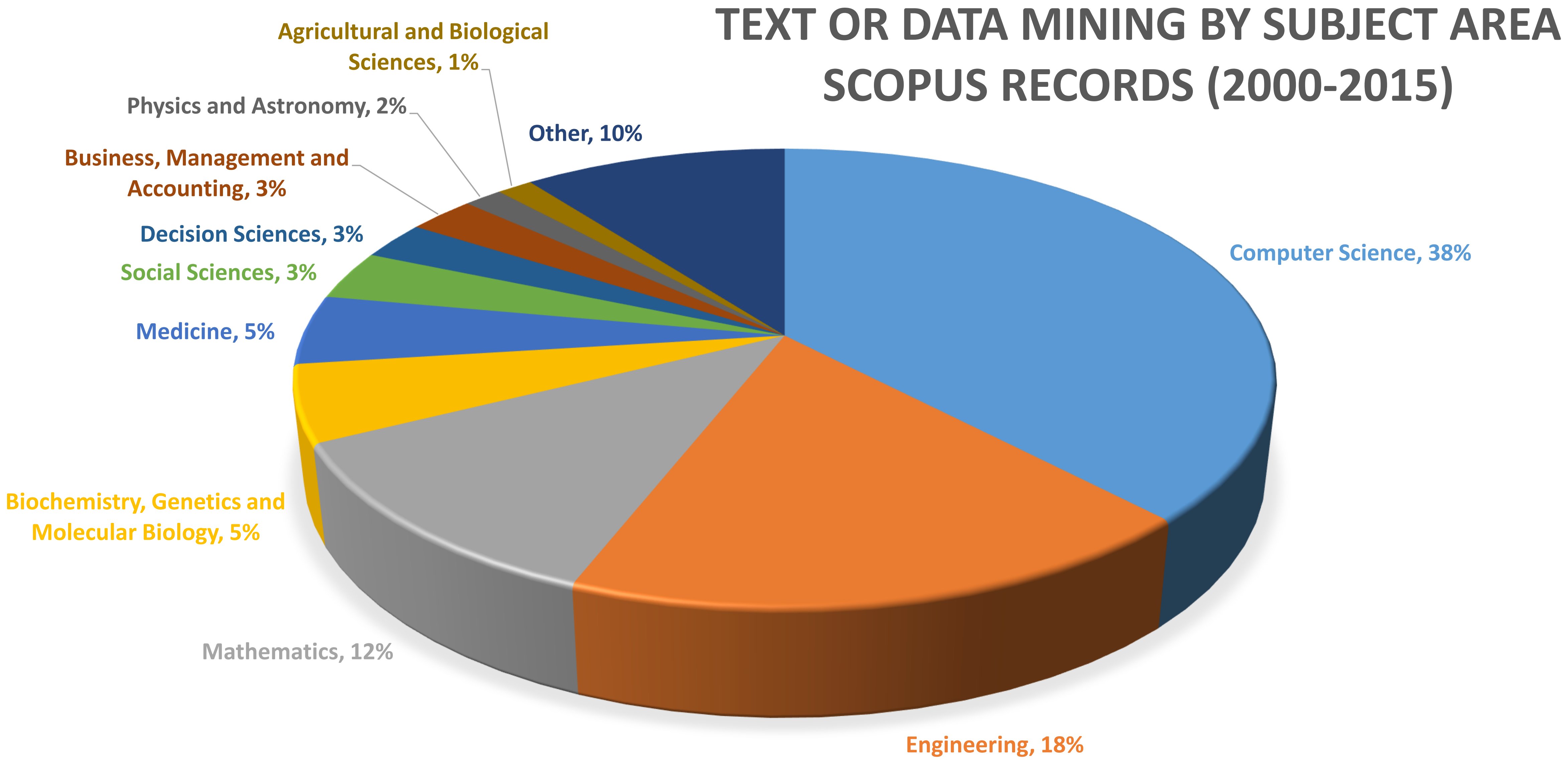
- "Text mining is arguably one of the most important fields driving growth, innovation, and even citizenship within a modern information economy."

PLOS Text and Data Mining Collection (Public Library of Science, n.d.)

- "Across all realms of the sciences, the rapid growth in the number of works published digitally presents new challenges and opportunities for making sense of this wealth of textual information."
- "The maturing field of Text Mining aims to solve problems concerning the retrieval, extraction and analysis of unstructured information in digital text, and revolutionize how scientists access and interpret data that might otherwise remain buried in the literature."

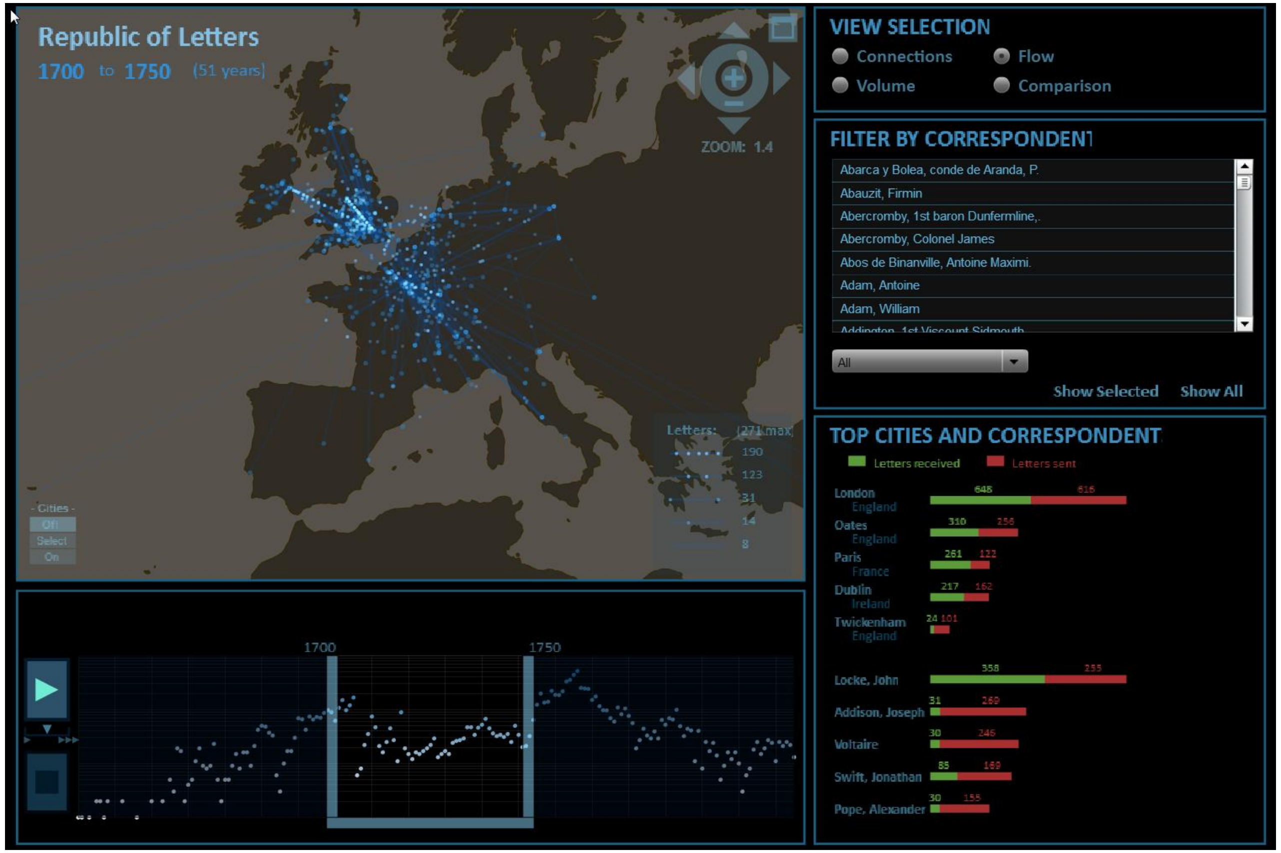
Text or Data Mining Records in Scopus by Year (54,719 total)





TEXT OR DATA MINING BY SUBJECT AREA SCOPUS RECORDS (2000-2015)

Mapping the Republic of Letters (Stanford)



Abarca y Bolea, conde de Aranda, P.
Abauzit, Firmin
Abercromby, 1st baron Dunfermline,.
Abercromby, Colonel James
Abos de Binanville, Antoine Maximi.
Adam, Antoine
Adam, William
Addington, 1st Viscount Sidmouth

	Letters received	Letters sent
London England	648	616
Oates England	310	256
Paris France	261	122
Dublin Ireland	217	162
Twickenham England	24	101
Locke, John	358	255
Addison, Joseph	31	269
Voltaire	30	246
Swift, Jonathan	85	169
Pope, Alexander	30	155

Center for Spatial and Textual Analysis (n.d.)

Potential Problems - Copyright

- You need datasets and many are copyright protected
- Getting permission for copyrighted data can be onerous and time-consuming
- Fair use or fair dealing rights are not clear and muddied by license language
- Often the negotiation falls on the researcher

Potential Problems - Technical

- Automated queries to vendor databases can overload their servers and cause headaches
- Often vendors prefer researchers to request a specific dataset to load locally on the researcher’s server, or they supply an application program interface (API)
- Data can take a long time to clean up to be usable

Role of Academic Libraries in TDM

- As argued by Orcutt (2015), “libraries do not recognize and, therefore, do not actively seek to support, emerging TDM interests among their researchers, leading those researchers to not view their libraries as a potential source for datasets”.

- Williams et al. (2014) believe that “librarians should assist in reducing transaction costs [for researchers] by developing model license clauses for text mining and routinely negotiating for these rights when the library purchases journals and other types of content”.

Research Questions

What percentage of the library's existing licenses include TDM clauses?

What percentage of the library's existing licenses have favourable TDM clauses?

Is there a difference in TDM terms between:

- consortial and local licenses?
- commercial and non-commercial publishers?

Methodology

Selected sample of 32 licenses by 30 different vendors/publishers

License selection based on:

- License availability (on file, up-to-date, signed)
- Content (quantity and suitability for TDM research)
- Publisher overlap (different product licenses by same vendor)
- Confidentiality clauses
- Representative samples of consortia and local licenses, different types of publishers, different disciplines/research areas, different material types

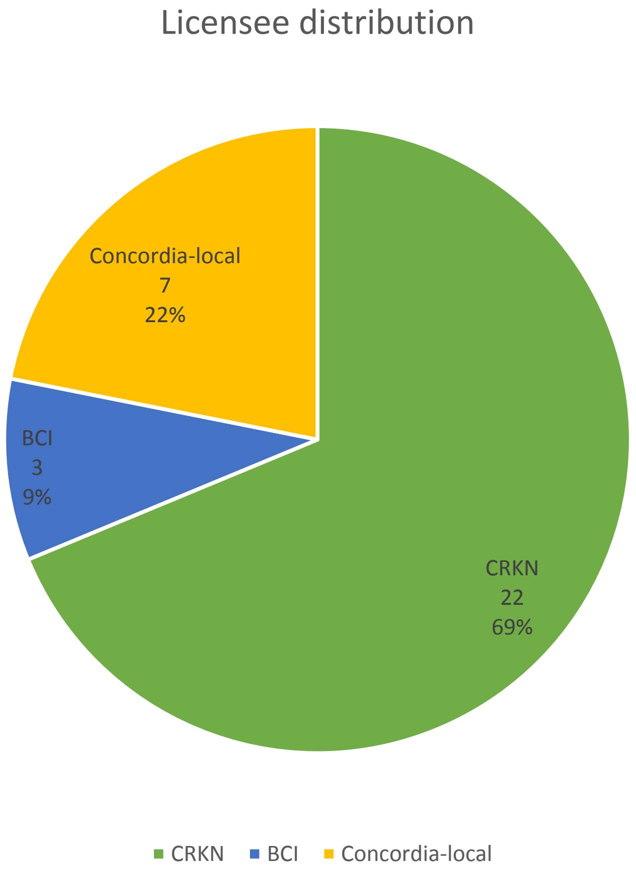
Excluded/not represented: eBooks, individual journals, streaming media, OA content

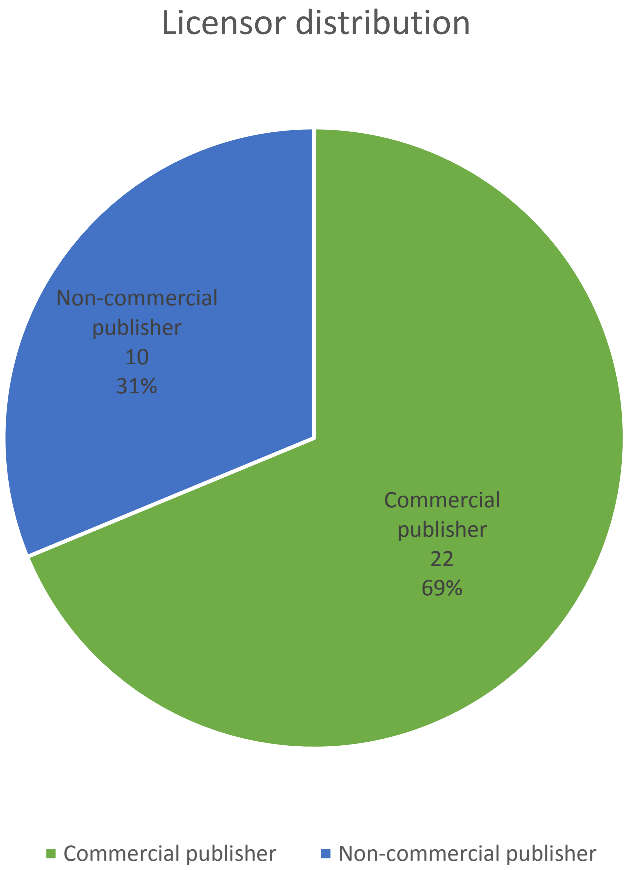
Methodology References: Beh and Smith (2012), Rogers (2009), Stemper and Barribeau (2006)

Licenses Sample

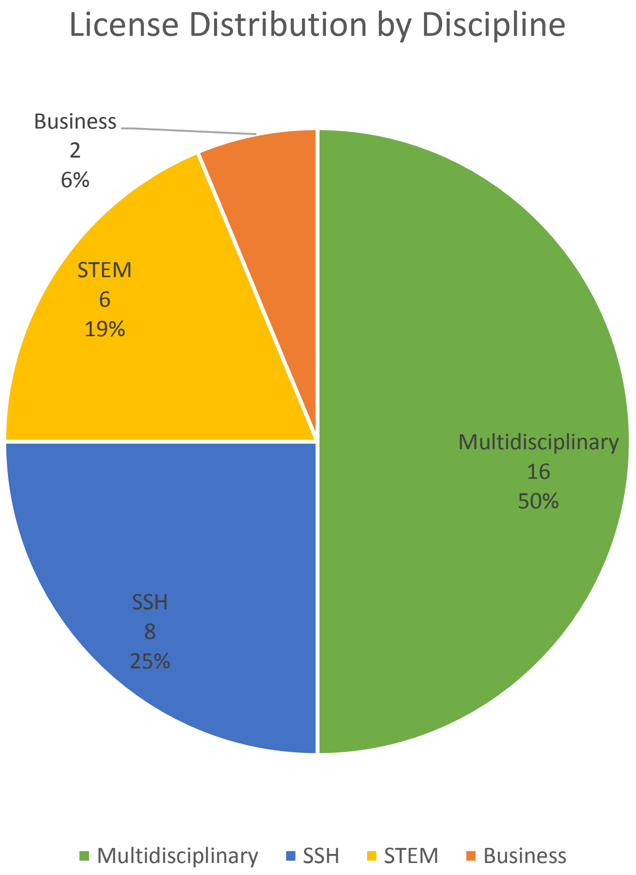
- ACM Digital Library
- American Chemical Society Web Editions & Legacy Archives
- Adam Matthew Digital
- Alexander Street Press, Various Databases
- American Mathematical Society, MathSciNet
- Annual Reviews
- ARTstor Digital Library
- CAIRN, Bouquet Général
- Canadiana.org, Early Canadiana Online
- Cambridge Journals Online
- EBSCO (various products)
- Elsevier ScienceDirect Journals
- Érudit, Cultural Magazines & Scholarly Journals Collection
- Gale Cengage Learning, ECCO and Times Digital Archive
- HeinOnline
- IBISWorld
- Institute of Physics Journals
- JSTOR Archive Collections
- Mergent Online
- NRC Research Press
- Oxford University Press Journals Online
- PhilPapers
- ProQuest (various products)
- ProQuest Canada's Heritage Globe & Mail 1844+
- ProQuest
- Royal Society of Chemistry Electronic Journals
- SAGE Publications
- SpringerLink Online Journals
- Taylor & Francis Journals Online
- Web of Science
- Vanderbilt Television News Archive
- Wiley-Blackwell Online Journals

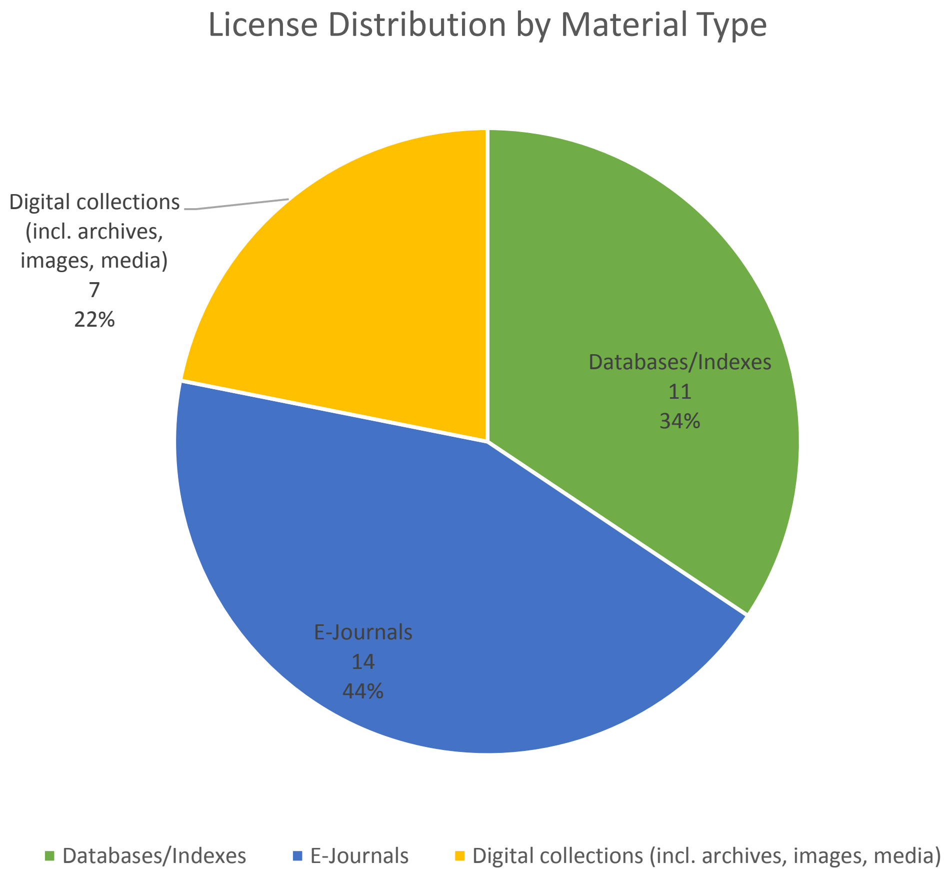
Distribution of Licenses Sample





Distribution of Licenses Sample





License Analysis

- Confidentiality clause
- TDM explicitly allowed
- License section to quote
- Any TDM “restrictions” (e.g. API only)
- Additional cost
- Additional agreement to sign
- Notes
- Additional TDM info (found outside license)
- Add later: Local loading clause

CRKN Standard TDM Clause

3. USAGE RIGHTS

3.1 The Consortium and the Members, subject to clause 6 below, may:

[…]

3.1.6 Apply automated tools and processes to the Licensed Materials, for the purposes of data mining for purposes of textual analysis and visual mapping of textual relationships, within the context of scholarship or research activities.

Modified CRKN TDM Clause

Example: American Chemical Society (2015)

3.1.6 Authorized Users and Walk-In Users may not modify, alter, or create derivative works of the materials contained in the Licensed Materials without prior written permission from ACS. With prior written permission from ACS, and subject to terms and conditions of a text and data mining addendum to be provided by ACS, Authorized Users and Walk-In Users may perform Text or Data Mining of Licensed Materials. […]

Silent on TDM / Related Usage Clauses

Example: Taylor and Francis (2007)

6 PROHIBITED USES

6.1 Neither the Consortium nor Authorized Users nor Walk-in Users may:

[…]

6.1.2 use any software such as webcrawlers, or any other means to systemically make print or electronic copies of multiple extracts of the Licensed Materials for any purpose.

Preliminary Results
TDM License Terms by Licensee

	CRKN		BCI		Concordia-local		TOTAL	
	N=22	%	N=3	%	N=7	%	N=32	%
Explicitly allow TDM	14	63.64	-	-	1	14.29	15	46.88
Explicitly forbid TDM	-	-	1	33.33	1	14.29	2	6.25
Do not mention TDM	8	36.36	2	66.67	5	71.43	15	46.88

Preliminary Results
TDM License Terms by Type of Publisher

	Commercial		Non-commercial		TOTAL	
	N=22	%	N=10	%	N=32	%
Explicitly allow TDM	9	40.91	6	60	15	46.88
Explicitly forbid TDM	2	9.09	-	-	2	6.25
Do not mention TDM	11	50	4	40	15	46.88

Preliminary Conclusions

- Favourable TDM terms in many existing library licenses
- CRKN-negotiated licenses most likely to allow TDM
- Licenses more likely to be silent than prohibit TDM
- Worth engaging publishers in negotiating TDM terms
- Promote availability of library resources for TDM research to campus community

Library Support for TDM

USC Libraries Research Guide:
http://libguides.usc.edu/textmining/databases

University of Chicago Library Guide:

http://guides.lib.uchicago.edu/textmining

MITLibraries / APIs for Scholarly Resources:

http://libguides.mit.edu/apis

Questions?

References

.txtLAB@McGill (n.d.). *Text mining the novel: A multi-university digital humanities initiative*. Retrieved from http://novel-tm.ca/?page_id=22

Beh, E., & Smith, J. (2012). Preserving the scholarly collection: An examination of the perpetual access clauses in the Texas A&M University Libraries' major e-journal licenses. *Serials Review, 38*(4), 235-242. http://dx.doi.org/10.1016/j.serrev.2012.10.005

Canadian Research Knowledge Network (2014, April 7). *Model license*. Retrieved from http://crkn.ca/programs/model-license

Center for Spatial and Textual Analysis (n.d.). *Mapping the Republic of Letters*. Retrieved from http://republicofletters.stanford.edu/

Orcutt, D. (2015). Library support for text and data mining. *Online Searcher, 39*(3), 27-30.

Public Library of Science (n.d.). *Text mining*. Retrieved from http://collections.plos.org/textmining

Rogers, S. (2009). Survey and analysis of electronic journal licenses for long-term access provisions in tertiary New Zealand academic libraries. *Serials Review, 35*(1), 3-15. http://dx.doi.org/10.1016/j.serrev.2008.11.002

Stemper, J., & Barribeau, S. (2006). Perpetual access to electronic journals: A survey of one academic research library's licenses. *Library Resources & Technical Services, 50*(2), 91-109.

Williams, L. A., Fox, L. M., Roeder, C., & Hunter, L. (2014). Negotiating a text mining license for faculty researchers. *Information Technology & Libraries, 33(3*), 5-21. http://dx.doi.org/10.6017/ital.v33i3.5485