

# DATA MINING RESEARCH WITH IN-COPYRIGHT AND USE-LIMITED TEXT DATASETS

A NATIONAL FORUM FUNDED BY THE INSTITUTE FOR MUSEUM AND LIBRARY SERVICES

## Project Team

Bertram Ludäscher (PI)

Beth Sandore Namachchivaya (Co-PI)

Megan Senseney (Co-PI)

Eleanor Dickson

MT Campbell

Lauryn Lehman

<https://publish.illinois.edu/limitedaccess-tdm/>

Twitter hashtag: [#TDMForum18](https://twitter.com/TDMForum18)



# IMLS NATIONAL FORUM APRIL 5 & 6, 2018 CHICAGO

LISTEN AND LEARN | SEEK COLLABORATIVE OPPORTUNITIES | MAKE COMMITMENTS



## GOAL


- Expand library-based **data services** to include provisions for supporting TDM with in-copyright and IP-restricted text data
- Develop recommendations for **best practices** and **policy** for text data mining services around:
  - Providing access to protected data
  - Documenting and disseminating research workflows for reproducibility
  - Hosting and preserving research outputs

# NATIONAL FORUM : APRIL 5-6, 2018

- Scott Althaus (University of Illinois at Urbana-Champaign)
- Christine Borgman (University of California, Los Angeles)
- Brandon Butler (University of Virginia)
- Beth Cate (Indiana University Bloomington)
- Marc Cormier (Gale-Cengage)
- Krista Cox (Association of Research Libraries)
- Mary Ellen Davis (Association of College and Research Libraries)
- J. Stephen Downie (University of Illinois at Urbana-Champaign)
- Patricia Feeney (Crossref)
- Lucie Guibault (Dalhousie University)
- Wolfram Horstmann (Göttingen University)
- Clifford Lynch (Coalition for Networked Information)
- Darby Orcutt (North Carolina State University)
- Thomas Padilla (University of Nevada, Las Vegas)
- Michelle Paolillo (Cornell University)
- Andrew Piper (McGill University)
- Peter Murray Rust (University of Cambridge)
- Matthew Sag (Loyola University, Chicago)
- Rachel Samberg (University of California, Berkeley)
- George Strawn (The National Academies of Science, Engineering, and Medicine)
- Jean Shipman (Elsevier)
- Paul Uhlir (independent consultant)
- Günter Waibel (California Digital Library)
- Kate Wittenberg (Portico, ITHAKA)
- Glen Worthey (Stanford University)

## RECOMMENDATIONS FOR LIBRARIES

- How, specifically, do we situate the library in this space?
- What do libraries need to know?
- What should libraries do?
- What do libraries need to do working in partnership with other groups?  
Who are they?
- What do other groups need to do? (e.g., what is outside the jurisdiction/  
auspices of the library?)



copyright law and resource licensing  
complicate research with text data



## *Text Data Mining (TDM)*

computational processes for applying structure to unstructured electronic texts and employing statistical methods to discover new information and reveal patterns in the processed data

---

*Use-Limited*  
~~*Limited Access*~~ *Data*

textual data where use and access are limited, or potentially limited, due to copyright, licensing, and other contractual terms



# METHODS

- Scoping review
- Participant identification
- SWOT analysis
- Forum statements
- “Liberating structures” for conversations at forum

# SWOT ANALYSIS

- Business models
- Content
- Legal & policy
- Library roles
- Publisher/content provider roles
- Research process
- Technical/technology

## W3 DEBRIEF (WHAT? SO WHAT? NOW WHAT?)

- TDM is part of a larger conversation:
  - *“This is about libraries making content useful and usable in the digital age”*
- More useful, usable content = accessible content
  - *“... make it about rights that we get - friendly conversation with people who have stuff we need”*
- Reading and content mining not mutually exclusive research activities
- Content mining can drive revenue if used appropriately

## MAKING COMMITMENTS

- Develop a declaration of principles
- Distill conversations into actionable recommendations for academic library services
- Outline the legal infrastructure for computational research with use-limited text datasets
- Write a grant proposal to develop legal and IP workshops for librarians and researchers
- Develop a pilot TDM service working with HathiTrust, Portico, publishers, and CrossRef
- Identify potential business models to support open content mining

## DEVELOP A DECLARATION ON TDM

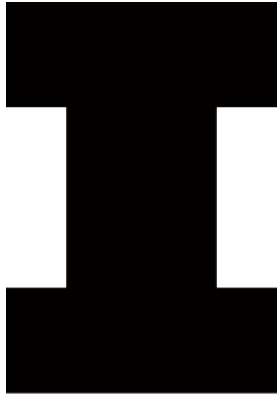
- Despite common efforts obstacles still remain at the national and international levels to the detriment of research and academic freedom
- Building on the movement toward open data in government and research, asserting that open data are minable data, by definition
- Building on the principles and recommendations put forward in The Hague Declaration on Knowledge Discovery in the Digital Age (2015) (<http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/> )
- Building on the FAIR principles for research data: Findable, Accessible, Interoperable, Reusable <http://dx.doi.org/10.1038/sdata.2016.18>

## NEXT STEPS

- White paper published by ACRL: Summer 2018
- Forum attendees working on individual and group commitments
- Want to get involved? Contact us: [bsnamach@uwaterloo.ca](mailto:bsnamach@uwaterloo.ca)

---

## PARTNERS



School of  
Information Sciences  
The iSchool at Illinois

 UNIVERSITY LIBRARY  
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



UNIVERSITY OF  
**WATERLOO**

## FUNDER



INSTITUTE *of*  
**Museum** and **Library**  
SERVICES

LG-73-17-0070-17