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

PRELIMINARY FINDINGS FROM A SYSTEMATIC LITERATURE REVIEW AND STAKEHOLDER INTERVIEWS

Megan Finn Senseney
University of Illinois
@modernmuchness

February 22, 2018

Senseney, M., Dickson, E., Namachchivaya, B. S., & Ludäscher, B. (2018). Data Mining Research with In-copyright and Use-limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews. Paper presented at the 13th International Digital Curation Conference, February 19-22, 2018, Barcelona, Spain.

copyright law and resource licensing complicate research with text data

SCENARIOS

3 REAL-WORLD EXAMPLES

#1 INCOMPATIBLE TERMS : **SCENARIO**

A researcher builds a dataset of text collected from an online collection, analyzes it, and submits a journal article based on the results.

#1 INCOMPATIBLE TERMS : **CONFLICT**

The journal returns reviewer comments and a request to share the data…

…but the Terms of Use disallow redistribution

> "Users are granted the right to download and store in machine readable form for 90 days, primarily for that User's exclusive use, a single copy of Material from _____."

#1 INCOMPATIBLE TERMS : **RESOLUTION**

The researcher negotiates to submit a supplemental report on methodological details, including:

- Details on data gathering process
- How data was extracted from raw text
- How data was processed for database
- The final database table structure

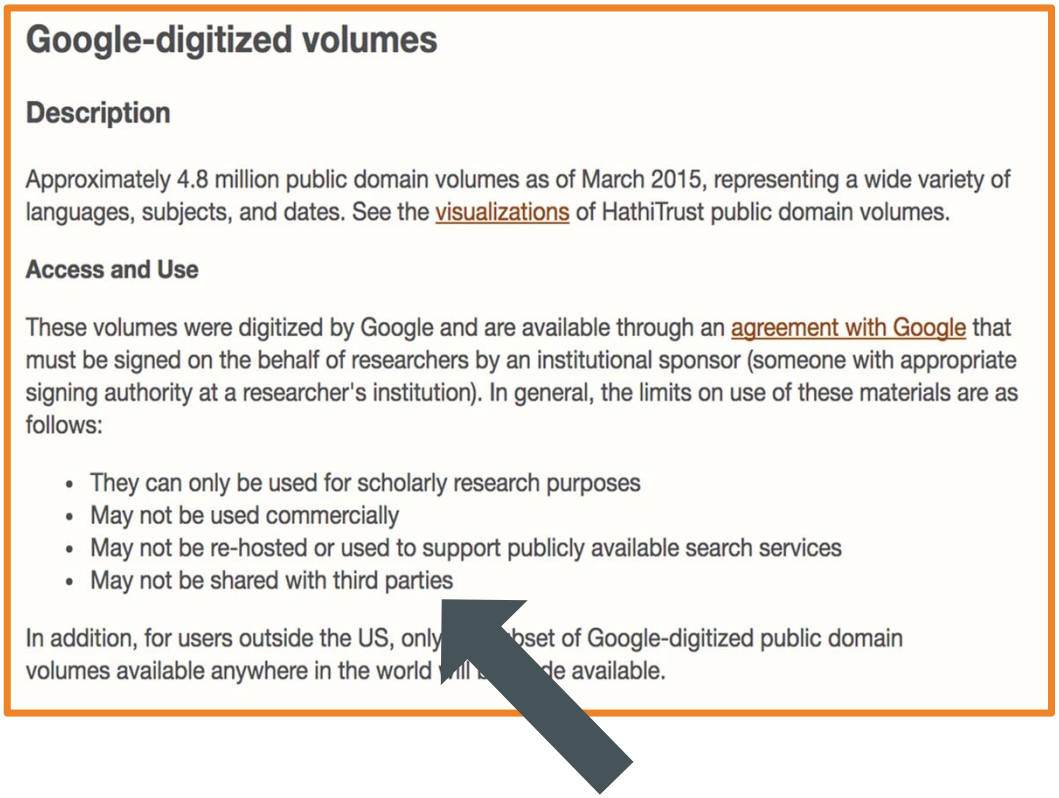
#2 DERIVED DATA : **SCENARIO**

- A researcher requests a custom dataset of works in public domain from HathiTrust
- Signs Google Agreement
- Acquires data
- Conducts research
- Submits paper to a scholarly journal

#2 DERIVED DATA : **CONFLICT**

- The researcher wants to share the data to support replication of the results…

  …but the Google Agreement disallows redistribution.



Google-digitized volumes

Description

Approximately 4.8 million public domain volumes as of March 2015, representing a wide variety of languages, subjects, and dates. See the visualizations of HathiTrust public domain volumes.

Access and Use

These volumes were digitized by Google and are available through an agreement with Google that must be signed on the behalf of researchers by an institutional sponsor (someone with appropriate signing authority at a researcher's institution). In general, the limits on use of these materials are as follows:

- They can only be used for scholarly research purposes
- May not be used commercially
- May not be re-hosted or used to support publicly available search services
- May not be shared with third parties

In addition, for users outside the US, only [illegible]bset of Google-digitized public domain volumes available anywhere in the world [illegible]de available.

#2 DERIVED DATA : **RESOLUTION**

- The researcher shares extracted word frequencies from the each volume in the dataset.
- The researcher posts these to a GitHub repository along with:
    - Python scripts
    - Bibliographic metadata
    - Derived results from analysis

#3 NOT-SO-PUBLIC DOMAIN : **SCENARIO**

A researcher who studies nineteenth-century fiction, wants to recreate a text analysis study he recently read. He begins to track down data for his own work.

#3 NOT-SO-PUBLIC DOMAIN : **CONFLICT**

He assumes it will be easy because all of the books analyzed are in the public domain…

> ...But soon discovers that public domain does not mean open. The data appears to have been drawn from multiple sources: open online collections, semi-open online collections, and a subscription database his university does not have access to.

#3 NOT-SO-PUBLIC DOMAIN : **RESULT**

The researcher abandons his project

:-(

SO WHAT DO WE DO?

- We're interested in how academic libraries can develop services to address these challenges
  - Within existing legal framework
  - Leveraging legal precedent that favors access
  - Thinking beyond access to documentation, sustainability, and reproducibility

BACKGROUND

ON TEXT DATA MINING AND USE-LIMITED 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

A NATIONAL CONVERSATION

IMLS National Forum

Text Data Mining with Use-Limited Datasets

~~Data Mining with In-Copyright and Limited-Access Text Datasets~~

July 1, 2017 – June 30, 2018

PI: Bertram Ludäscher

Co-PIs: Beth Namachchivaya and Megan Senseney

Co-Investigator: Eleanor Dickson

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 Horstman (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)
- Gunter Waibel (California Digital Library)
- Kate Wittenberg (Portico, ITHAKA)
- Glen Worthey (Stanford University)

GOAL

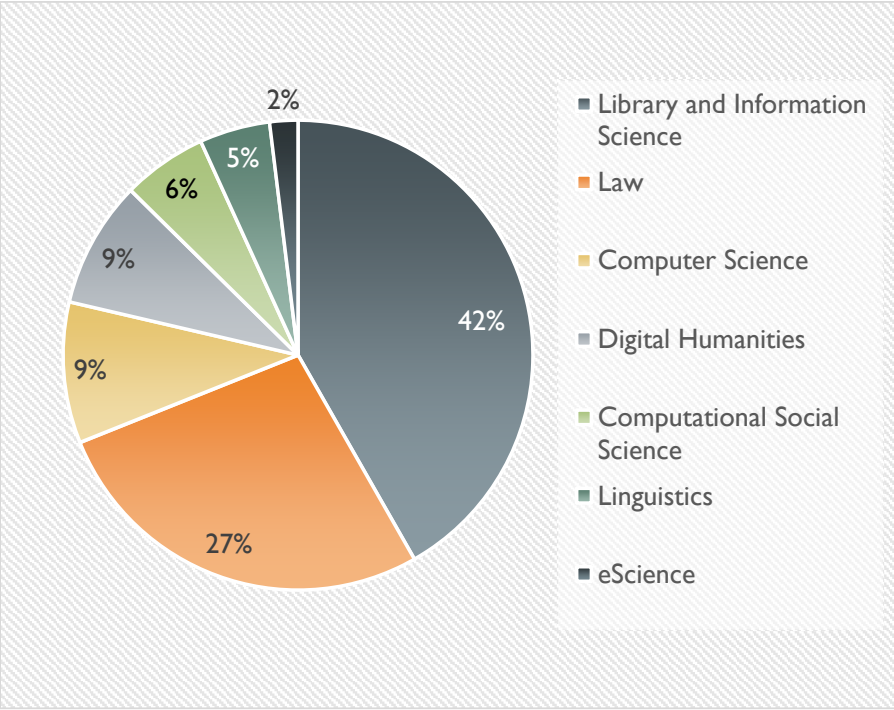
- Expand library-based data services to include provisions for supporting TDM with in-copyright and IP-restricted text data
- Develop best practices and policy for text data mining services around:
  - Providing access to protected data
  - Documenting and disseminating research workflows for reproducibility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

METHOD

LITERATURE REVIEW | STAKEHOLDER INTERVIEWS | FORUM STATEMENTS | SWOT ANALYSES

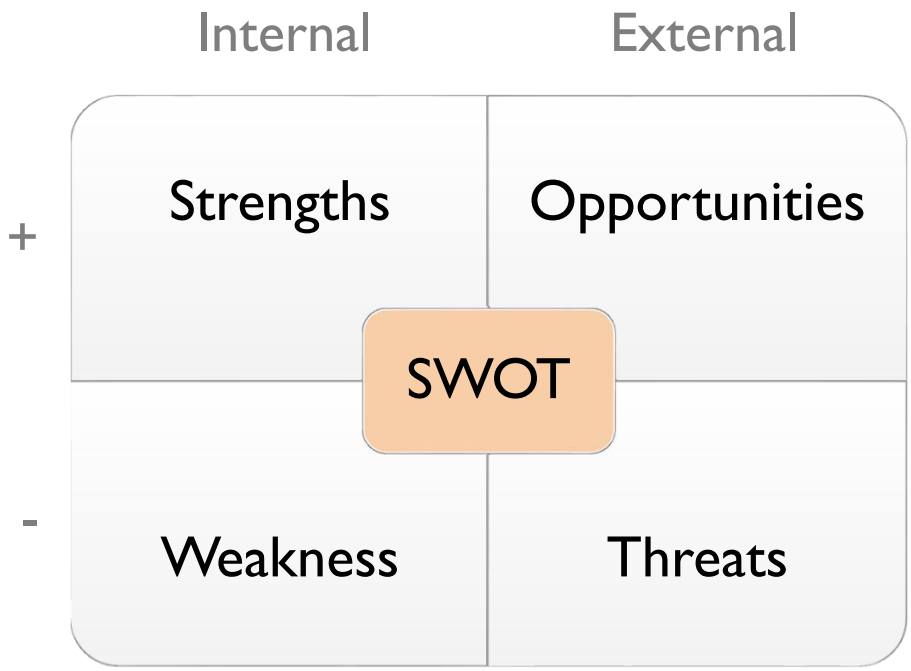
LITERATURE REVIEW

- Initial search: 103 results across 7 domains
- Citation chaining: 40 additional items
- Compiling
  - Tools
  - Case Law
  - Model Licensing Agreements
- Goal: Conduct environmental scan and integrate prior literature into emerging discussions



INTERVIEWS

- 25 semi-structured interviews
- Introduce SWOT structure and discuss initial ideas for forum statement
- Identifying
  - Key themes
  - Relationships among themes
  - Points of convergence and divergence across stakeholder groups
- Goal: Develop strategies for managing risk



PRELIMINARY FINDINGS

LEGAL UNCERTAINTIES AND LEGAL BOUNDARIES

- Best practices for fair use in TDM
- License negotiations to gain access
- Institutional and jurisdictional boundaries
- Legal action as opportunity or threat

On fair use and licensing in data mining:

"There's very solid legal ground to say that text and data mining is a fair use. I think there's a very strong legal argument to be made that unless your license explicitly denies you the right to rely on fair use that you should work under the assumption that you can rely on fair use."

POLICY AND ADVOCACY

- Institutional risk management
- Ongoing cross-stakeholder engagement
- Awareness building and policy development with professional societies and government agencies
- Advocacy for copyright exemptions and open access policies.

On strategizing for advocacy:

"This is an area that requires a lot of collective focus and a lot of collective action. It's not anything any individual organization or university can tackle by itself. So I feel like the strategy of working on this nationally is a good one."

TRAINING AND SUPPORT

- Scholarly communication
- Data literacy
- Introductory training
- Professional re-framing and re-skilling
- Scaling up

On data selection:

"I became aware of how many of our researchers were doing mining work, and they were pretty much just grabbing any old dataset."

STANDARDIZATION AND ACCESS WORKFLOWS

- Shared terminology
- Data transfer
- Data formats
- Data quality

On standardizing data across vendors:

"Obtaining the text in any kind of standardized form is, I would say, quite impossible. When you obtain text from a vendor-supplied database, it comes however the vendor decides to give it to you."

NEXT STEPS

TOWARD RECOMMENDATIONS AND BEST PRACTICES

A MULTI-PRONGED APPROACH

Local / Institutional

- Integrate TDM into library-based digital scholarship services to
  - Consolidate knowledge and lessons learned across units
  - Establish standardized local procedures
- Support services should address strategies for:
  - Evaluating consumptive vs. non-consumptive approaches to TDM,
  - Assessing data quality,
  - Acquiring and integrating datasets as needed,
  - Processing and analyzing data,
  - Documenting research workflows,
  - Packaging results in shareable formats to support reproducibility, and
  - Developing strategies for communicating results.

Global / Cooperative

- Articulate and distribute action across stakeholder groups to
  - Formalize more uniform agreements and data transfer practices
  - Establish best practices and disciplinary norms
  - Advocate for legislation that enacts a copyright exception or codifies more open policies for text data mining
  - Participate in multi-national copyright harmonization efforts
  - Develop international standards for interoperability and data interchange
  - Build a business case for vendor and publisher engagement

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