Digging Deeper into Text and Data Mining

By Inga Haugen, Edward F. Lener, Virginia Pannabecker, & Philip Young

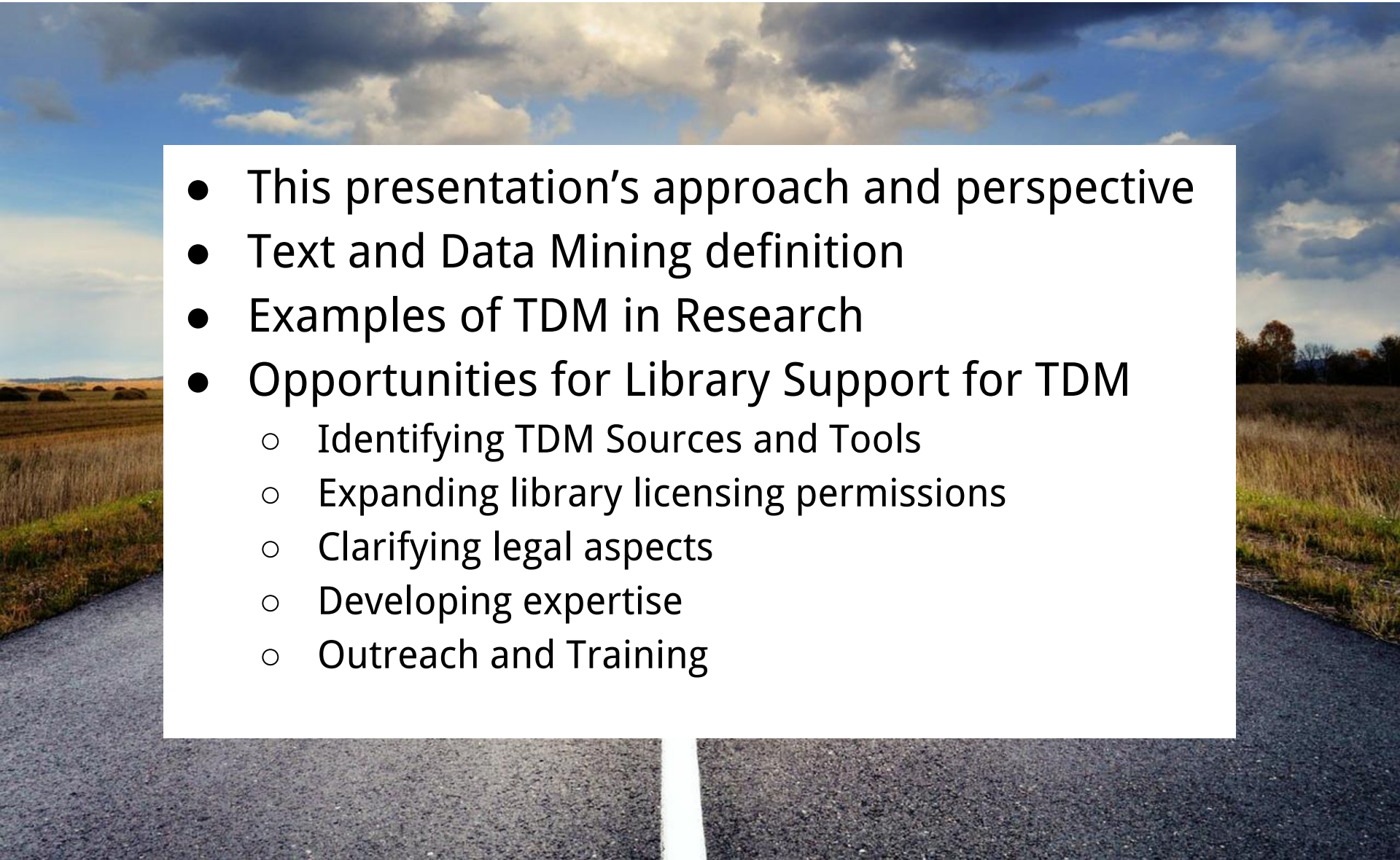
Virginia Tech, University Libraries

Presentation available in VTechWorks institutional repository

http://hdl.handle.net/10919/79483

VIRGINIA TECH™

Roadmap / Outline



- This presentation’s approach and perspective
- Text and Data Mining definition
- Examples of TDM in Research
- Opportunities for Library Support for TDM
  - Identifying TDM Sources and Tools
  - Expanding library licensing permissions
  - Clarifying legal aspects
  - Developing expertise
  - Outreach and Training

Introductions - Presenters

Inga Haugen Agriculture, Life Sciences, and Scholarly Communication Librarian; Interim liaison to the College of Natural Resources and the Environment (CNRE)

Edward Lener Associate Director for Collection Management and College Librarian for the Sciences

Ginny Pannabecker Associate Director for Research Collaboration and Engagement; Liaison Librarian for life sciences and biomedical programs

Philip Young Institutional Repository Manager


Buna ziua
Dzień dobry
SZIA
SELAM
HEJ
HEJSAN
SVEIKI
GUTEN TAG
ALIO
NOROC
SANNU
Tjänare
SERVUS
Cześć
CIAO
BONJOUR
PRONTO
HELLO
SZERVUSZ
BONGHJORNU
DAR FIA
Olá
Allô
HALOO
Hi
DIA DUIT
SALVE
TIENS
ALOHA
ZDRAVO
MERHABA
VERWELKOMING
Hylô
HALLO
HEI
SALUT
ZDRAVO

What is Text and Data Mining (TDM)?

Text and data mining (TDM) uses methods of automated extraction, combination, and analysis of data to create new information by revealing trends, patterns, and relationships. The mining of text and of data usually require different considerations. Text mining, sometimes called text analytics, can be viewed as a subset of data mining.

Examples of TDM in research / scholarship

From VT

TDM Forum - variety of topics and examples - business, cyber bullying, statistics

History - 1918 influenza pandemic project

Other examples

Health

Discovering associations between adverse events from electronic health record data

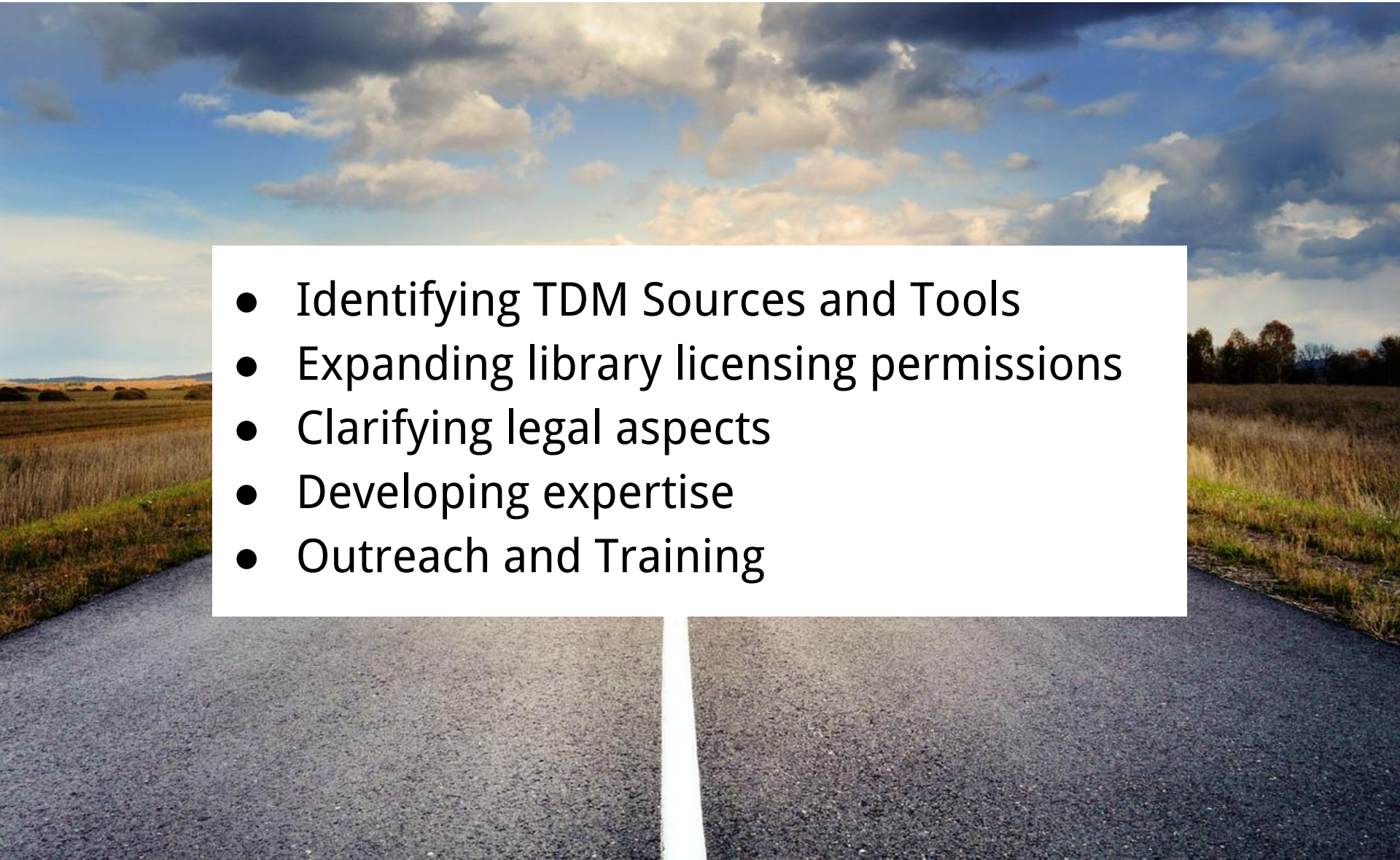
Digital Humanities

Librarian collaborations for research and training in text encoding

Opioid Crisis

A Text Mining Analysis of Public Reactions to Opioid Crisis

TDM: Academic library support opportunities



- Identifying TDM Sources and Tools
- Expanding library licensing permissions
- Clarifying legal aspects
- Developing expertise
- Outreach and Training

Examples of academic library support for TDM

Identifying & Sharing TDM Sources and Tools

- Tool and Methods Guides
  - UC Berkeley Library guide
  - MIT Libraries - APIs guide
  - University of Melbourne - Text Mining Tools list
  - Carnegie Mellon University Libraries guide
  - VT Libraries
- Conversations with researchers / Community of practice
- Open educational training options
- Conducting literature reviews for updates to to stay current in best practices and tools

Examples of academic library support for TDM

Expanding library licensing permissions

- Each license represents an opportunity
- Vendors are at widely different places on this issue
- Liblicense model agreement and others like it can be a very helpful starting point
- Need to decide how important this issue is to you and what you are prepared to accept

Examples of academic library support for TDM

Clarifying legal aspects*

- Subject to U.S. law
- No specific exemption for TDM in U.S. copyright law**
- License agreement overrides Fair Use provisions of copyright law
- Researchers may need data from multiple sources with varying rights

ARL Issue Brief
Text and Data Mining and Fair Use in the United States

* Note - We are not attorneys!

**You may hear about a specific tdm exemption in the UK - more info here.

Examples of academic library support for TDM

Sample license language from one major library vendor

> *"TDM Output" means the result of any TDM activity by Researcher, such as the creation of an index, abstract, relative or absolute description or representation of the Content; any algorithm, metrics, method, standard or taxonomy describing or based on the Content, … whether in a the form of a direct extraction or a representation in any form which is based on any portion of the Content. Any quotes from the Content shall be limited to fifty (50) words or less.*

Examples of academic library support for TDM

Developing expertise

- Join Communities
  - Text and Data Mining Research Support List (JISC)*
- Read / Start a Journal Club - How-To Texts, Research studies that used text and data mining methods
  - Introduction to Text Analysis
- Tutorials / Training / Webinars
  - Coursera Data Science course via Johns Hopkins
  - Hathi Trust Research Center | Training
  - Programming Historian
  - R and Data Mining Courses - Directory of free options
  - Software Carpentry / Data Carpentry / Library Carpentry

*JISC (formerly Joint Information Systems Committee) is a UK non-profit supporting higher education

Examples of academic library support for TDM

Outreach and Training

Host events / Workshops / Discussions

- Open Data Day
  - Invite a local Code for America brigade to co-host or facilitate a hack-a-thon or workshop
- ContentMine
  - Identify researchers or applications producers you'd like to work with and work with them to provide an event.
- TDM Forum
  - Invite researchers in your community/institution to share their experiences and expertise

Employ TDM to improve library services

'Bibliomining'

Examples

"Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries."

"Use and understand: the inclusion of services against texts in library catalogs and "discovery systems"

"Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts"

"Data-mining the Library"

Activity
Consulting on a TDM Project: Questions to Ask

If you knew someone were coming to you for a 'TDM Consultation,' **what questions might you ask?**:

- Ahead of time (if you have the opportunity)
- At the start of the consultation
- During the consultation (what are key info points you'd want to be sure to touch on during the conversation)

Activity
Consulting on a TDM Project: Questions to Ask - Examples

- Does the content source provider support TDM projects, provide policies, or offer contacts for requests?
- How will the file type or extraction method affect the TDM project goals?
- How large is the expected data? How will it be transferred? Where will it be stored?
- If the data is digitized text, what is the quality of the Optical Character Recognition (OCR)? Is it possible to get a sample of the OCR to check for quality?

→ **What would you add?**

Additional References

Report: Young, P., Brittle, C., Haugen, I., Lener, E., Pannabecker, V. (2017). **Library support for text and data mining: A report for the University Libraries at Virginia Tech**. Retrieved from http://hdl.handle.net/10919/78466

This Presentation: http://hdl.handle.net/10919/79483

Libraries, licensing, and TDM

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

Organization to Follow

- Future TDM

Selected recent items of interest since our report was shared

- Text mining of 15 million full-text scientific articles
- JSTOR Labs - Text Analyzer for Topics and Recommended Readings
- Legal Analytics Lab at Georgia State University
- Common Crawl web archiving organization | Related article

Images

Sourced from Pixabay - CC0 - public domain license

https://pixabay.com/en/binary-binary-system-data-dataset-2728121/
https://pixabay.com/en/road-asphalt-space-sky-clouds-220058/
https://pixabay.com/en/balloon-discussion-comment-2223048/
https://pixabay.com/en/hello-bonjour-hi-greeting-foreign-1502369/
https://pixabay.com/en/chat-multiple-icon-symbol-message-2389223/
https://pixabay.com/en/contract-consultation-pen-signature-1332817/
https://pixabay.com/en/weight-scale-equal-arm-balance-scale-2402966/
https://pixabay.com/en/home-office-workstation-office-336377/
https://pixabay.com/en/people-girls-women-students-2557396/
https://pixabay.com/en/silhouette-head-bookshelf-know-1632912/
https://pixabay.com/en/banner-header-question-mark-1090830/
https://pixabay.com/en/icon-feedback-message-cloud-data-1968237/

Questions?

Evaluation Link

Please let us (and VLA) know your thoughts on this presentation!

tinyurl.com/th2017vla

Thank You word cloud by Ashashyou [CC BY-SA 4.0], via Wikimedia Commons

VIRGINIA TECH™