Operationalization, MOU Automation, and Just Where Exactly is the Data: Stories about Overcoming Access Issues for Text and Data Mining Data

Colloquium On Text & Data Mining in Libraries 2023
Sam Hansen Mathematics & Statistics Librarian
University of Michigan, Ann Arbor
https://tinyurl.com/tdmLibColl2023

We own a lot of data, but…

We own a lot of data, but…

1. The format it comes in can be hard to use

We own a lot of data, but…

1. The format it comes in can be hard to use
2. Some of our licenses require user agreements

We own a lot of data, but…

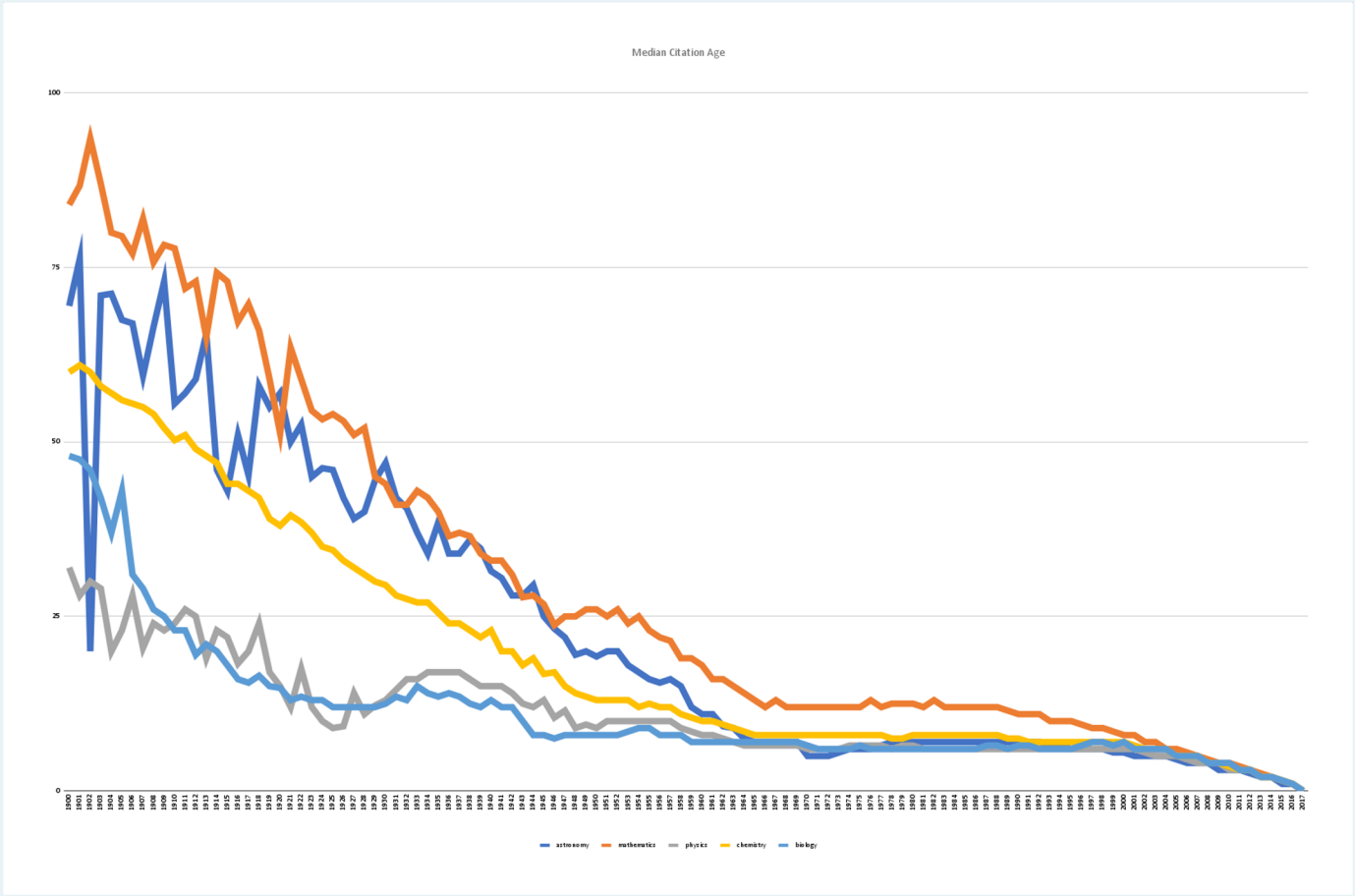
1. The format it comes in can be hard to use
2. Some of our licenses require user agreements
3. We do not even know what and where all the data is

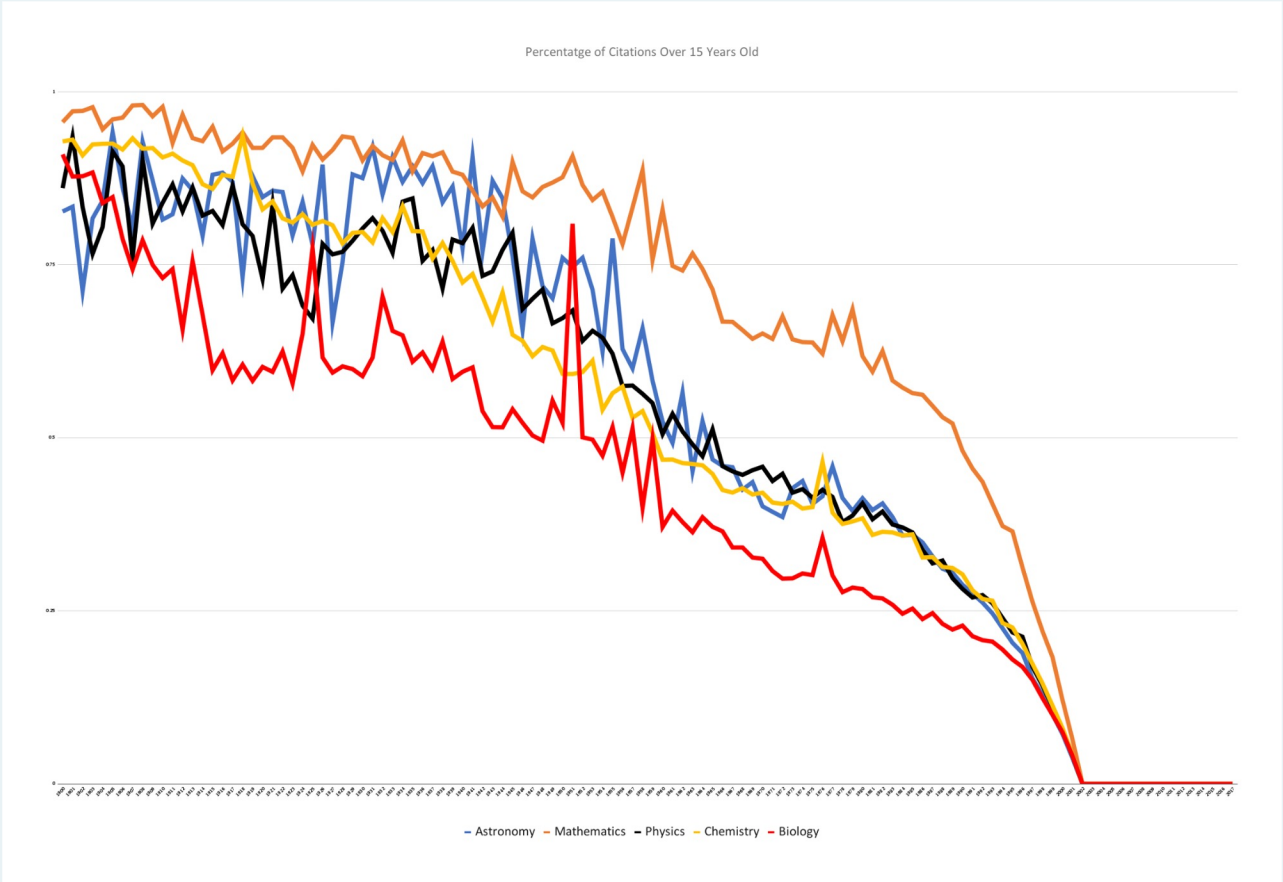
Problem Solving and Service Development

- Two Main TDM Groups at U-M Ann Arbor Libraries
  - TDM Advisory Group
  - Digital Scholarship Advisory Group subcommittee - TDM Service Design Team
- Our Three Focus Projects
  - Operationalizing Web of Science XML
  - Automating Memorandums of Understanding
  - Identifying Licensed TDM Data Sources

WoS Operationalization

- Clarivate provides XML
  - It is a headache to work with
- We are transforming the XML into three different formats
  - DBGZ
  - TSV
  - SQL
- Many thanks to wonderful people at University of Toronto and University of Indiana for the code which allowed us to accomplish this!
- RIP CADRE


Median Citation Age
100
75
50
25
0
astronomy
mathematics
physics
chemistry
biology


Percentatge of Citations Over 15 Years Old
0.75
0.5
0.25
0
Astronomy
Mathematics
Physics
Chemistry
Biology

MOU Automation

- We have been gathering user MOUs for datasets via email, hard copy, and digital forms
- Created a pilot Qualtrics form which included authentication and display logic for our two most requested datasets
- Still need to automate providing access once a MOU is signed

M | LIBRARY

Please choose the data set which you need access to:

- ◯ Clarivate Web of Science
- ◯ CoreLogic

Submit

University of Michigan Library
913 S. University Avenue
Ann Arbor, MI 48109
(734) 764-0400

Additional Conditions of Use
You agree to:

- Communicate the conditions outlined in this MOU to any and all additional researchers, research assistants, and other involved personnel who will have access to the digital content files, to make sure they are aware of and agree to abide by these conditions.
- Include an acknowledgement (in a foreword, introduction or footnote) in any publications based on research conducted with these digital content files that the files are licensed material obtained with the assistance of the University of Michigan Libraries

Do you agree to abide by these additional terms:

☐ Yes, I agree with the Additional Conditions of Use

Please briefly describe your research inquiry or purpose with this data set or digital files in 100 words or less:

Effective Dates:
This MOU is effective upon signature and remains in effect until amended as agreed upon by both Parties. I have read, understand, and agree to abide by the terms outlined in this agreement.

Please type your name to indicate your agreement:

Finding (and Organizing) the Data

- Asked Selectors if they are the gatekeepers of any datasets, and if so how they provide access
  - Add user to online service
  - Add user to shared drive
  - Add new account role to user
  - Provide login for service

Finding (and Organizing) the Data

- We also dug into the various places where datasets are stored
  - Vendor Platform
  - Dropbox
  - Google Drive
  - Shared Network Drives
  - Physical Media
    - Flash drives, hard drives, CDs
      - Not always in catalog
  - A librarian’s desktop computer

Finding (and Organizing) the Data

- Created, and still populating, an internal spreadsheet of datasets
- Exploring capabilities of new LMS to better catalog data
- Continuing our investigation into what datasets have been purchased and/or licensed

	A	B	C	D	E	F	G	H
1	**This is for Library staff. This isn't exhaustive, please don't share the link to this editable spreadsheet with patrons.**							
2	**Title of Resource**	**Publisher/Creator**	**Data Locally Held**	**Data Location**	**Data Size**	**Locally Held Data Details**	**API Available?**	**Data Available via API**
3	American Israelite	ProQuest	1854-1925	Library MiStorage		Article-level metadata (1 text file for each article)		
4	BBC Monitoring	BBC	None				Yes	
5	Boston Globe	ProQuest	1872-1983	Library MiStorage		Article-level metadata (1 text file for each article)		
6	Chicago Defender	ProQuest	1909-1975	Library MiStorage		Article-level metadata (1 text file + 1 PDF for each article)		
7	Chicago Tribune	ProQuest	1849-1935	Library MiStorage		Article-level metadata (1 text file + 1 PDF for each article)		
8	CoreLogic Parcel Level Real Estate Data	Corelogic	1990 or earlier - 2016	Library MiStorage	Deed – 35 GB Foreclosure – 6 GB Tax – 12 GB	Parcel level data on buildings/homes, 2016 housing tax data and foreclosure		no
9	Crossref Metadata API	Crossref	None				Yes	Metadata of every DOI for 4000+ Crossref publisher members
10	Detroit Free Press	ProQuest	1831-1922	Library MiStorage		Article-level metadata (1 text file for each article)		
11	Detroit Free Press	ProQuest	1923-1999	Library MiStorage		Page-level metadata only (1 text file for each full page)		
12	Dimensions Plus	Digital Science	None				Yes	

¿Questions?