Metadata as data

OhioDIG: Leveraging Repositories for Digital Projects II

Nick Ver Steegh

2020-07-23

Hello

- I'm Nick
- Ohio University
- Metadata Services Department
- Metadata Technologies Librarian

Metadata Technologies Librarian

- Collaborate on Digital Initiatives projects
- Bibliographic MARC cataloging
- Specialization in technology
  - OpenRefine, XML, macros etc.

Presentation

1. Data & Data Science
2. Metadata Science
3. Collection metadata dashboard prototype
4. Reflection

Data

- Metadata is data
- Data can be studied
- Therefore metadata can be studied

Example: R's `mtcars`

	mpg	cyl	disp
Mazda RX4	21.0	6	160.0
Mazda RX4 Wag	21.0	6	160.0
Datsun 710	22.8	4	108.0
Hornet 4 Drive	21.4	6	258.0

`https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html`

Uses of data

- Academic research (e.g. bioinformatics, social sciences)
- Government (e.g. census, *1984*)
- Business (e.g. Equifax, Facebook)

Data Science

- Programming + Domain expertise + Statistics
- Workflow:
    1. Find data
    2. Wrangle data
    3. Explore data
    4. Model data
    5. Publish/deploy model

Techniques of Data Science

- Programming
- Data visualization
- Statistics
- Machine learning

Tools of Data Science

- Proprietary software suites
  - MATLAB, SPSS
  - Tableau
  - Excel
- Open-source programming languages
  - R
  - Python

Summary

Data scientists analyze and model numerical data by doing statistics, machine learning, and data visualization using software tools like Excel and Python.

Data in libraries

- We provide data (data librarianship)
- We store data (research data management)
- We analyze data (collection/operations assessment)
- We create data (cataloging, digital collection metadata)

The special case of metadata

- On the levels of measurement Library metadata is mostly *nominal*
- There are few statistical tools for nominal data
  - Names can be `=`, i.e. counted
  - Operations like `>`, `<`, `+`, `-`, `*`, `/` are invalid
- Dublin Core (DC) supports
  - Faceting
  - Keyword search
  - Basic authority control

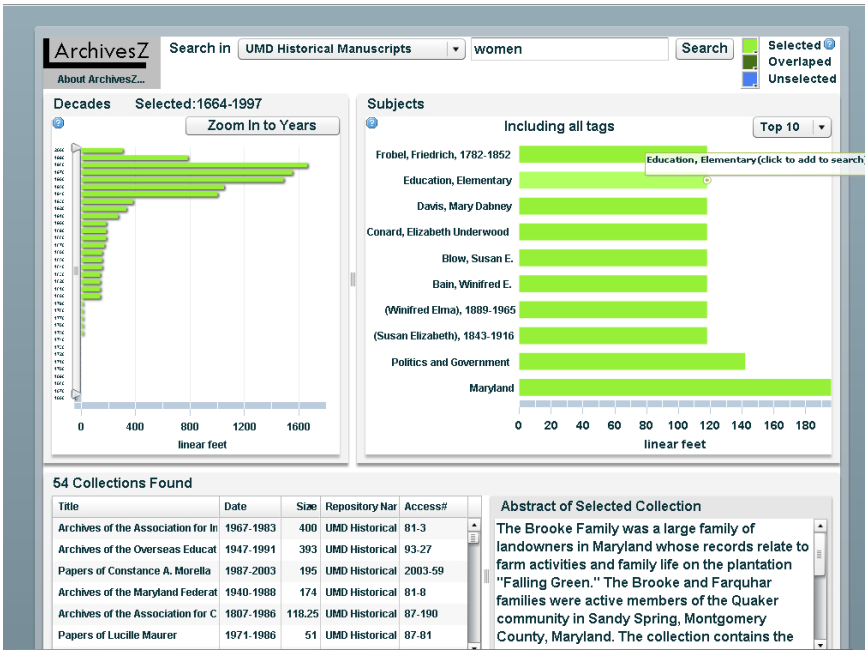
Why metadata as data?

- New form of access
  - Now: reading room access
  - Now: online (point-and-click) access
  - *New*: data access
- New forms of discovery
  - Now: Keyword search
  - Now: Controlled vocabularies
  - *New*: timelines, maps, network graphs etc.
- Reciprocal benefit
  - Analysis benefits access
  - Access benefits analysis

Prior work

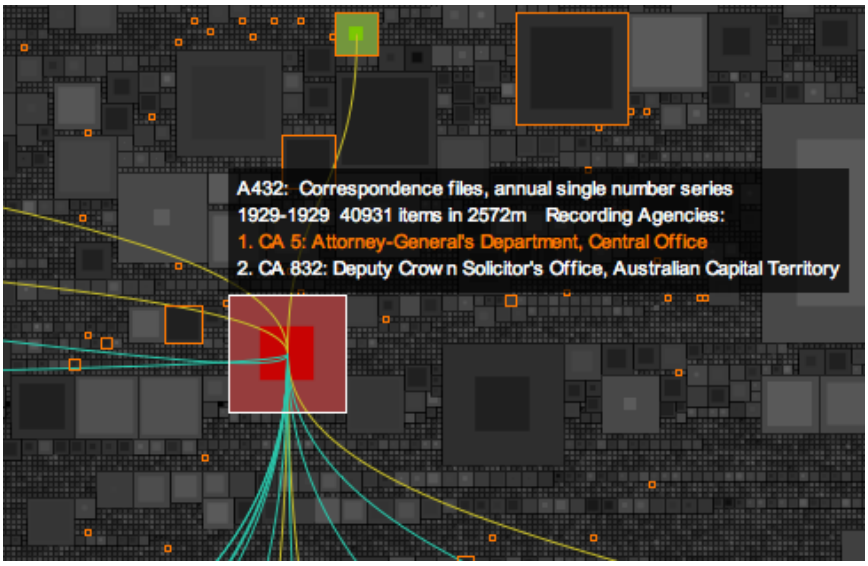
- Bahde, Anne. "Conceptual Data Visualization in Archival Finding Aids: Preliminary User Responses." Portal: Libraries and the Academy, vol. 17, no. 3, 2017, pp. 485-506. DOI.org (Crossref), `doi:10.1353/pla.2017.0031`.

ArchiveZ (2007)



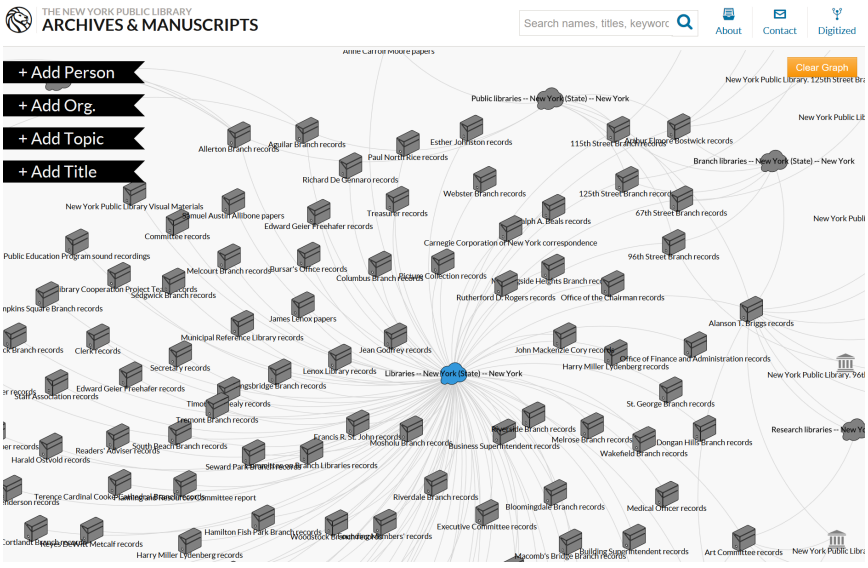
Kramer-Smyth, Nishigaki, Anglade. "ArchivesZ: Visualizing Archival Collections"

The Visible Archive (2009)



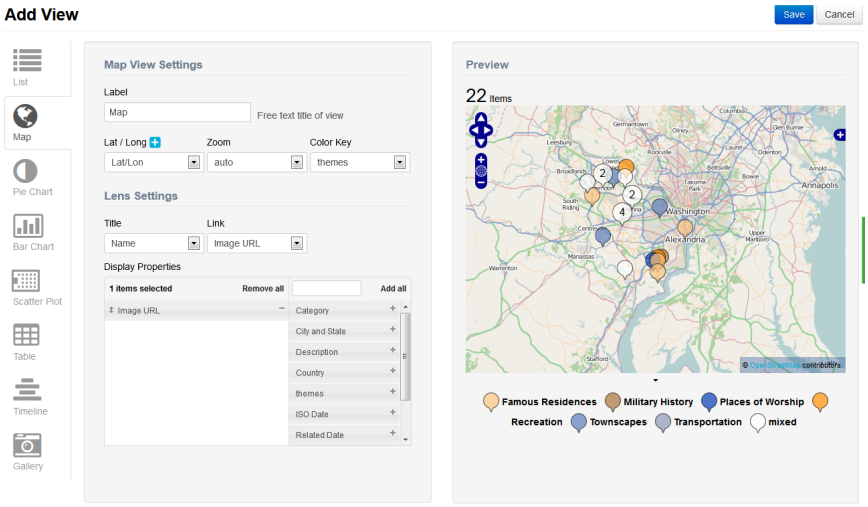
http://visiblearchive.blogspot.com/

NYPL Terms Explorer (2013)



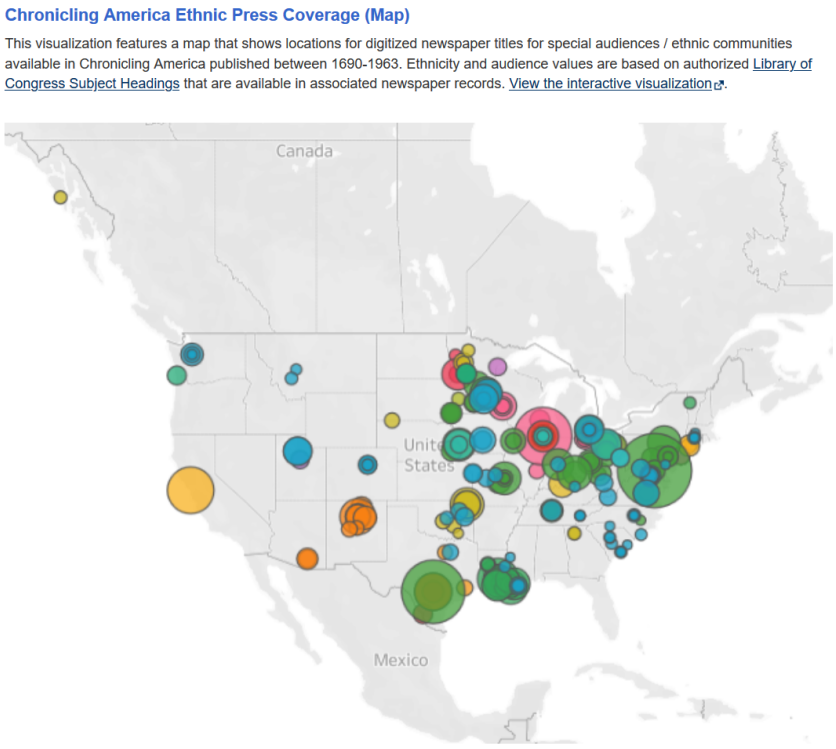
http://archives.nypl.org/terms/

ViewShare (2011)



https://blogs.loc.gov/thesignal/category/viewshare-2/

Chronicling America (2019)



`https://www.loc.gov/ndnp/data-visualizations/`

Metadata science

- Could we be getting more out of our metadata? With science?
- Data Science as a service of Metadata Services?
- Experiment: Build a single-page dashboard for browsers

Visualizations of interest

- Maps
- Timelines
  - Histograms of dates

- Network graphs
- Bar and/or pie charts
- Interactivity for engagement

Scope of interest

- Repository
- Collection
- Item

The technology

- Applications (Tableau, Excel, ViewShare R.I.P.)
  - Lower learning curve
  - Less flexible
  - Expensive
- Programming languages (R, Python)
  - Higher learning curve
  - Flexible
  - Free as in beer

Programming languages

- R: great for statistics
- Python: the second best language at everything
- JavaScript: required for the web

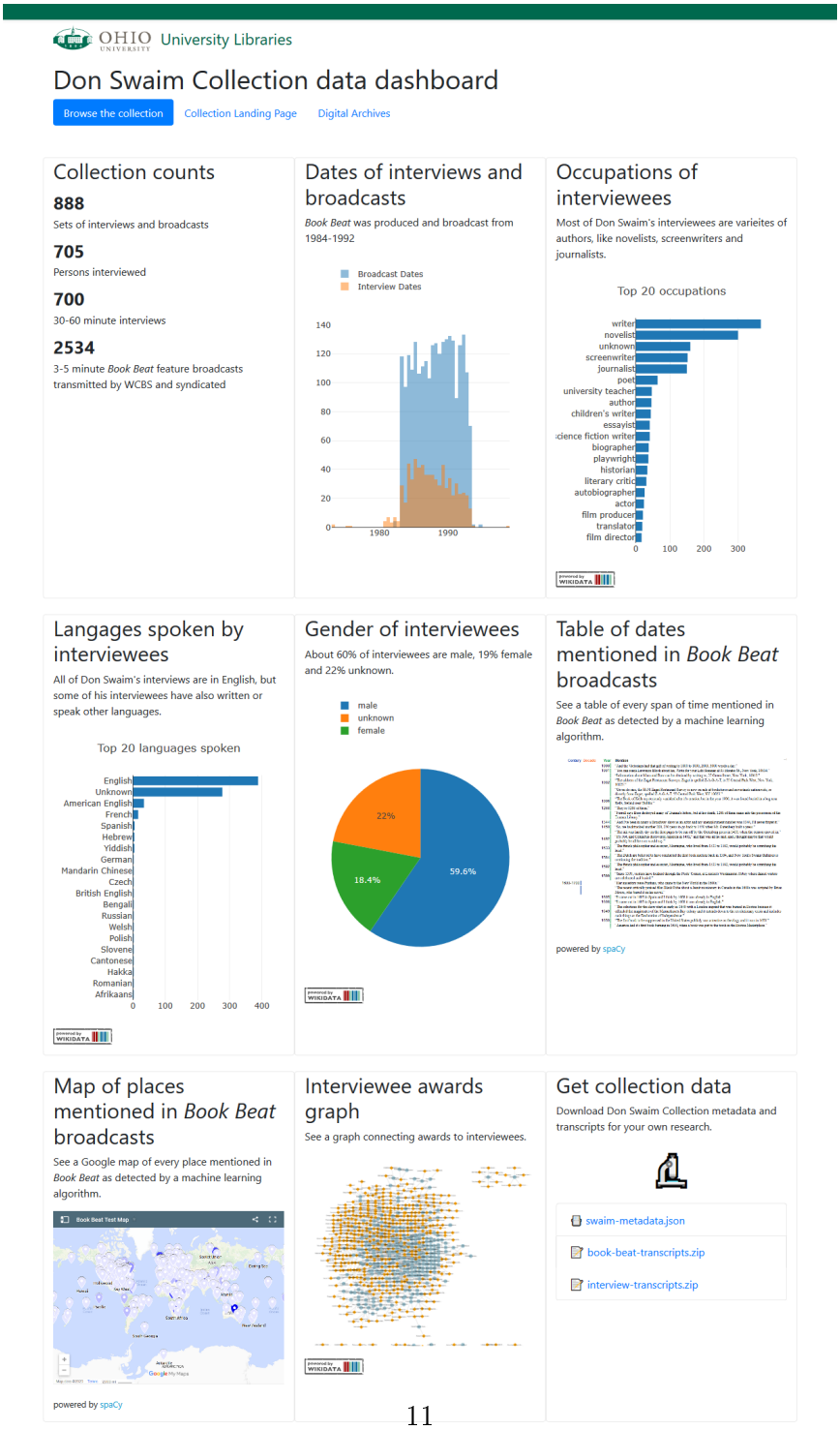
The workflow

1. Python: get data from CONTENTdm, Wikidata
2. Python: wrangle and explore the data
3. HTML5+CSS+JavaScript: build the dashboard

JavaScript visualization libraries

- Hundreds
- I chose:
  - Plotly.js for charts
  - Cytoscape.js for network graphs
  - Timeline.js for timelines
- Alternatives to JavaScript
  - R's Shiny
  - Python's Dash, Bokeh, Streamlit, HoloViz

Dashboard demo screenshot



Reflections

- More work, for perhaps only marginal benefit to researchers
- Lukewarm support from DC and contemporary DCMSs
- Muddles the purpose of metadata (support DCMS *and* Data Science)
- Blurs the boundary between librarianship and digital humanities
- Sustainability: how much technology do we want to learn, as a profession? As institutions?
- Finally a use case for the Semantic Web?

The future

- OCLC Research's CONTENTdm Linked Data Pilot
- Librarians & Archivists x Vendors

Thanks!