Grow Your Role: The Rare Book Data Project
An approach to learning collections through data

Jennifer King
Leah Richardson
George Washington University
MARAC Spring 2018
The Rare Book Data Team:

Research and User Services:
Dolsy Smith, Humanities Librarian + Collections Strategist

Special Collections:
Jen King, Manuscripts Librarian
Leah Richardson, Research + Outreach Librarian

Goals:

- Collaborative learning experience with colleagues
- Learn more about our rare book holdings to inform teaching and collection development
- Intro to Python programming language
- Inform our skill and ability to support Collections as Data researchers
Collections as Data

At head of title: Leaves of Grass.

Author’s name, Walt Whitman, appears in copyright statement on title page verso.

Reissued, with half-title, in Two rivulets (1876). Cf. Myerson.

Preface -- One song, America, before I go -- Souvenirs of democracy -- As a strong on bird on pinions free -- The mystic trumpeter -- O star of France! -- Virginia--the west -- By broad Potomac’s shore.
How do archivists support this type of research?

How do we provide access to the data?

Should librarians and archivists provide training on how to use data and specific tools?

Should libraries create the tools?

Workshops offered in our library
Look at MARC data for rare book holdings as a large scale dataset

Answer questions about the size, breadth, depth, coverage, strength, gaps of our holdings

Questions we wanted to ask:

- How many rare books do we have?
- What is the range of publication dates?
- Where are the places of publication?
- Who were the printers?
- Who were the publishers?
- What languages are represented?
- What are the subjects?

Phase 1: What do we want to learn from the data?
What is the range of publication dates?
Where are the places of publication?
Who were the printers?
Who were the publishers?
What languages are represented?
What are the subjects?

Bound in vellum. [5 DGW | 9 LOCAL

Casuistry | v Early works to 1800.

Conscience | v Early works to 1800. [0

Casuistry. | 2 fast | 0 http://id.worldcat.org

Conscience. | 2 fast | 0 http://id.worldcat.org

Early works. | 2 fast | 0 http://id.worldcat.org
● 73,000+ records in initial data extraction

● Dolsy organized MARC data into a .csv file using PyMARC
  ○ Data parsed as:
    Row for each Bib#, column for each field and subfield

Phase 2: Select MARC Fields
Phase 3: Data Extraction
Used Python to do most data manipulation; saved work in Jupyter Notebook environment
  ● Learned Python in a task-based approach

Tasks:
  ○ Replace column names with plain language
  ○ Normalize dates
  ○ Exclude MS call numbers

ASB student working with OpenRefine:
  ○ Cleaned up the 650 field
  ○ Removed trailing punctuation
  ○ Removed unicode characters that didn’t translate

Subfields posed a challenge
  ○ Data parsed as column for each field + subfield
  ○ BUT subfields interact with each other to create meaningful details about items--working to capture these connections in next iteration

Phase 4: Data Clean-up (and more data clean-up)
What did we learn?

About the collection:
- Number of books?...We can tell you the number of unique bib IDs: 60,238
- Date coverage: 1470s to 2017
- Almost immediately we were better situated to enhance support for rare book instruction and collection development
- Including MARC in rare book instruction
- Rare book collections as University Archives; this data allowed us to see the work of librarians, archivists, and curators and place ourselves in that genealogy
- Rare book collections as University Archives: Engaging with the research around GW’s history with slavery can be supported through the collecting history

Enhanced professional skills:
- We learned a lot about MARC--its potential and limitations
- We learned a lot about Python and ourselves

Phase 5: Querying the Cleaned Data
Number of Rare Books per Decade 1470-2010

● Use visualizations to aid discovery
● Build a short-title catalog
● HathiTrust API:
  ○ Using OCLC identifiers
  ○ What do we have access to?
  ○ Explore our data with HTRC tools
● Promote our collections for their computational research value
● Make our dataset and code available online
● Ethical considerations around Collections as Data?
  ○ Privacy and primacy
  ○ This is ONE type of research, we serve many communities

Next steps
- Look at metadata and consider it; every MARC record and finding aid is a work of scholarship
- Think about how rare book and archives users talk about access:
  - I’m looking for books printed in England during the early part of the 17th century
  - I’m interested in the Shaw neighborhood of DC during the 1970s
- Regular Expression or “regex”
- Accessing data: APIs, ILS; read documentation
- Other tools: OpenRefine, HTRC
- Consider the knowledge and expertise all around you: your colleagues

Getting started with Collections as Data
Thank you

Contact us:

Leah Richardson
leahr@gwu.edu
Twitter: @rareleah

Jen King
jenking@gwu.edu
Twitter: @jenkingarchives

Special Collections @ GW:
go.gwu.edu/archives

Twitter: @gwuspeccoll