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**

029 1_ |**a** NZ1 |**b** 5604138
035 __ |**a** (OCoLC)1035105 |**z** (OCoLC)19230731 |**z** (OCoLC)42776723
035 __ |**a** (OCoLC)ocm01035105
040 __ |**a** DLC |**b** eng |**e** rda |**c** DLC |**d** UBY |**d** OCLCQ |**d** OCLCG |**d** OCLCA |**d** OCLCO |**d** OCLCA |**d** OCLCQ |**d** DGW
049 __ |**a** DGWW
050 00 |**a** PS3207 |**b** .A1 1872
100 1_ |**a** Whitman, Walt, |**d** 1819-1892. |**0** http://id.loc.gov/authorities/names/n79081476
245 10 |**a** As a strong bird on pinions free, and other poems.
264 _1 |**a** Washington, D.C. : |**b** [Walt Whitman?], |**c** 1872.
264 _3 |**a** (New-York : |**b** S.W. Green)
300 __ |**a** x, [16], 8 pages ; |**c** 21 cm
336 __ |**a** text |**b** txt |**2** rdacontent |**0** http://id.loc.gov/vocabulary/contentTypes/txt
337 __ |**a** unmediated |**b** n |**2** rdamedia |**0** http://id.loc.gov/vocabulary/mediaTypes/n
338 __ |**a** volume |**b** nc |**2** rdacarrier |**0** http://id.loc.gov/vocabulary/carriers/nc
500 __ |**a** Advertisement, "Walt Whitman's books," on last 8 pages.
500 __ |**a** At head of title: Leaves of grass.
500 __ |**a** Author's name, Walt Whitman, appears in copyright statement on title page verso.
500 __ |**a** Reissued, with half-title, in Two rivulets (1876). Cf. Myerson.
505 0_ |**a** 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.
510 4_ |**a** BAL |**c** 21408
510 4_ |**a** Myerson, J. Whitman, |**c** pages 183-187
590 __ |**a** Inscribed on front flyleaf in black ink: "Josiah Royce from Walt Whitman the author Dec: 7 '90." |**9** LOCAL

Collections as Data

HACK-TO-LEARN

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?

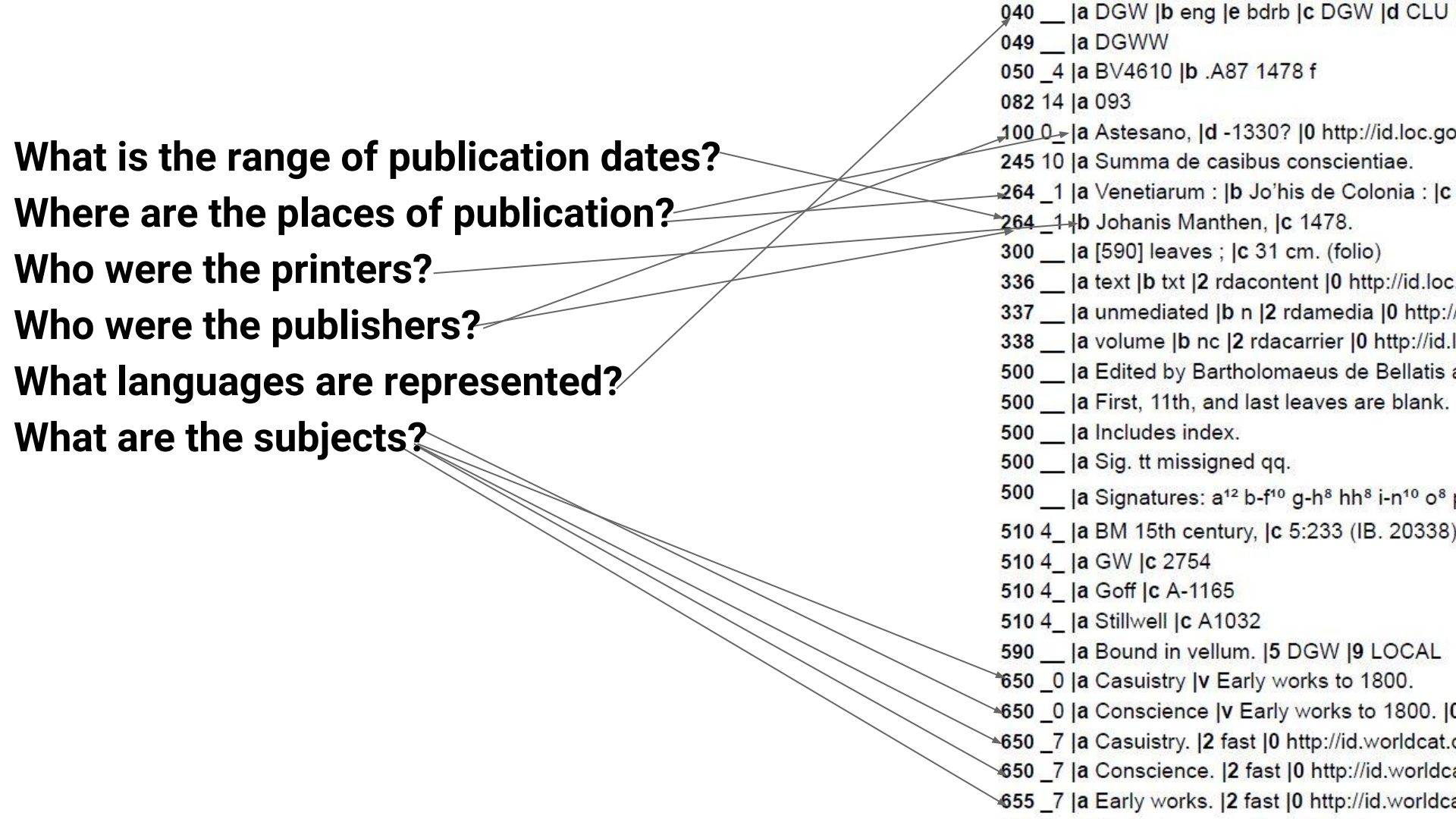
Where are the places of publication?

Who were the printers?

Who were the publishers?

What languages are represented?

What are the subjects?



040 __ |a DGW |b eng |e bdrb |c DGW |d CLU
049 __ |a DGWW
050 _4 |a BV4610 |b .A87 1478 f
082 14 |a 093
100 0_ |a Astesano, |d -1330? |0 http://id.loc.go
245 10 |a Summa de casibus conscientiae.
264 _1 |a Venetiarum : |b Jo'his de Colonia : |c
264 _1 |b Johanis Manthen, |c 1478.
300 __ |a [590] leaves ; |c 31 cm. (folio)
336 __ |a text |b txt |2 rdacontent |0 http://id.loc
337 __ |a unmediated |b n |2 rdamedia |0 http://
338 __ |a volume |b nc |2 rdacarrier |0 http://id.l
500 __ |a Edited by Bartholomaeus de Bellatis a
500 __ |a First, 11th, and last leaves are blank.
500 __ |a Includes index.
500 __ |a Sig. tt missigned qq.
500 __ |a Signatures: a^{12} b-f^{10} g-h^{8} hh^{8} i-n^{10} o^{8} p
510 4_ |a BM 15th century, |c 5:233 (IB. 20338)
510 4_ |a GW |c 2754
510 4_ |a Goff |c A-1165
510 4_ |a Stillwell |c A1032
590 __ |a Bound in vellum. |5 DGW |9 LOCAL
650 _0 |a Casuistry |v Early works to 1800.
650 _0 |a Conscience |v Early works to 1800. |0
650 _7 |a Casuistry. |2 fast |0 http://id.worldcat.o
650 _7 |a Conscience. |2 fast |0 http://id.worldca
655 _7 |a Early works. |2 fast |0 http://id.worldca

001: Control number
008: Publication date
041: Language
100: Author
245: Title
260: Publication information
264: More publication information
300: Physical description
500: Notes
541: Local note (gift statement)
6xx: Subjects
7xx: Provenance

- 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

Phase 4: Data Clean-up (and more data clean-up)

- **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**

In [ ]:
```
new_columns = {'041-1_-a': 'translation_languages',
               '041-0_-a': 'additional_languages',
               '041-1_-h': 'original_language',
               '041-__-a': 'additional_languages',
               '041-__-h': 'original_language',
               '100-0_-a': 'mono_author_name',
               '100-10-a': 'author_name',
               '100-1_-a': 'author_name',
               '100-20-a': 'author_name',
                '100-2_-a': 'author_name',
               '100-3_-a': 'author_name_family',
               '752-__': 'added_place',
               '300-__-a': 'extent',
               '300-__-b': 'other_physical_details',
               '300-__-c': 'dimensions'
              }
```

In [23]:
```
column_map = {}
for k,v in new_columns.items():
    if v in column_map:
        column_map[v].append(k)
    else:
        column_map[v] = [k]
```

In [5]:
```
title = {'a': 'title',
        'b': 'remainder of title',
        'c': 'statement of responsibility'}
```

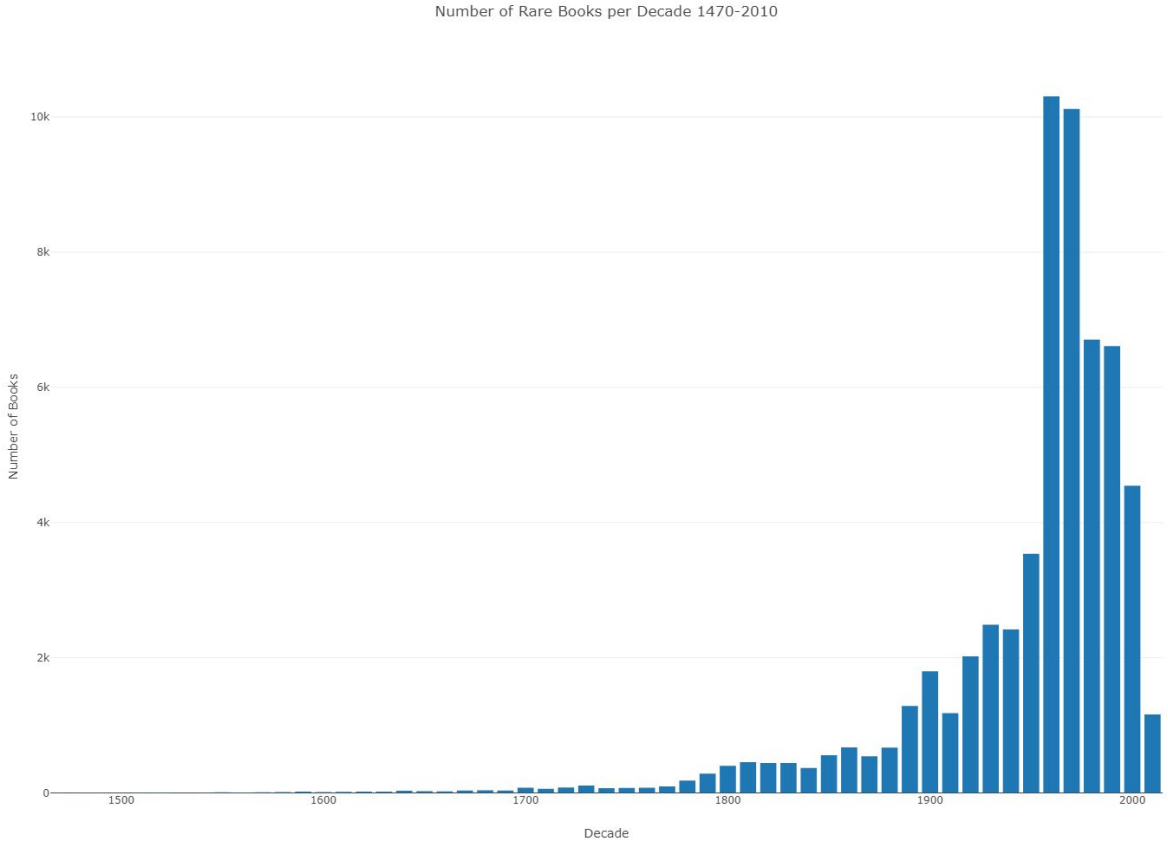
What did we learn?

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

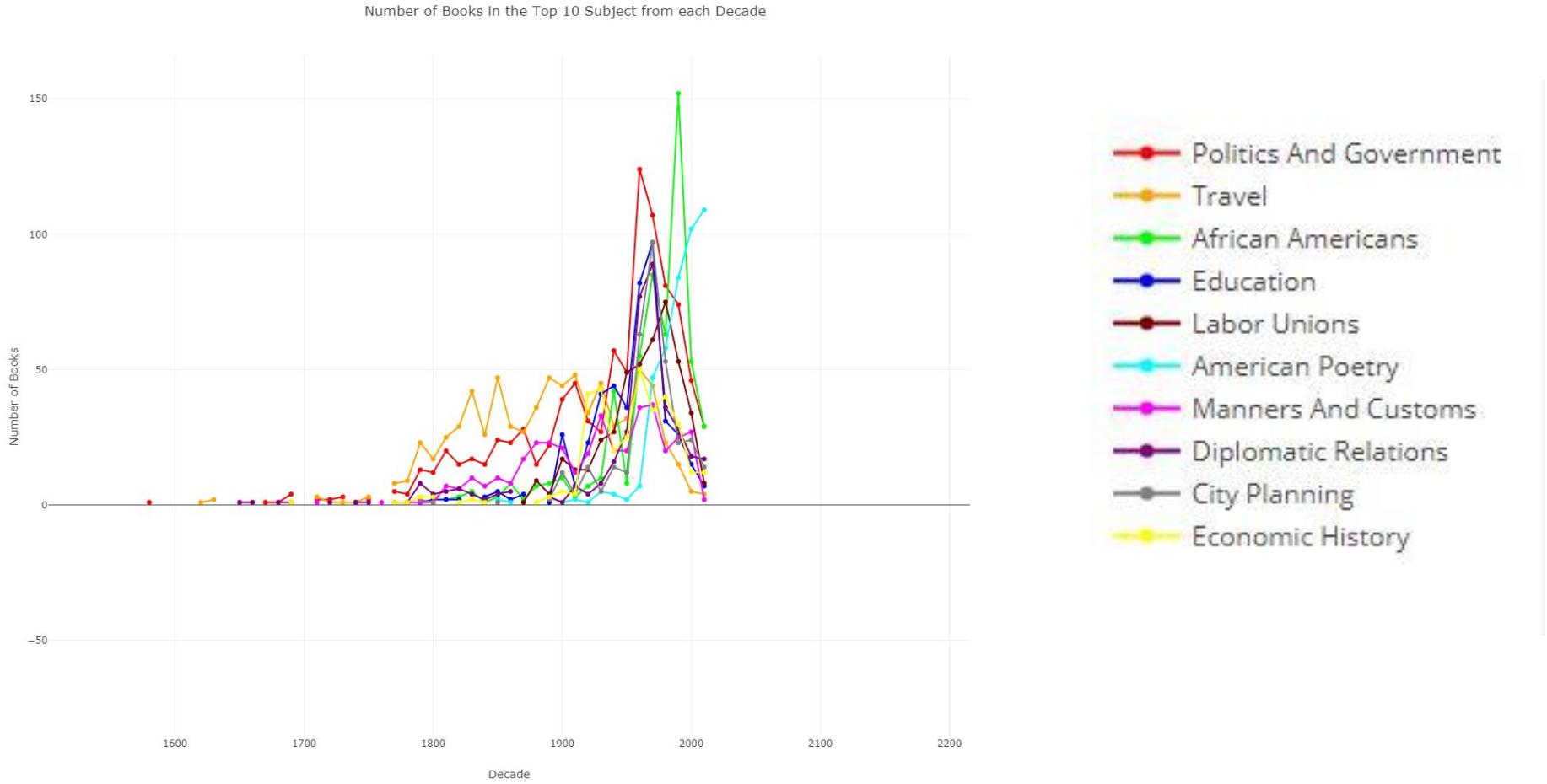
Enhanced professional skills:

- We learned a lot about MARC--its potential and limitations
- We learned a lot about Python and ourselves



https://plot.ly/~ktopham/26/number-of-rare-books-per-decade-1470-2010/

x	y		
1470	1		
1480	1	1780	183
1490	3	1790	283
1500	1	1800	399
1510	4	1810	455
1520	4	1820	441
1540	3	1830	442
1550	8	1840	369
1560	4	1850	556
1570	10	1860	672
1580	12	1870	542
1590	19	1880	669
1600	13	1890	1285
1610	15	1900	1800
1620	18	1910	1181
1630	19	1920	2019
1640	32	1930	2487
1650	25	1940	2419
1660	23	1950	3539
1670	33	1960	10305
1680	38	1970	10121
1690	34	1980	6706
1700	75	1990	6609
1710	61	2000	4543
1720	80	2010	1160
1730	109		
1740	71		
1750	72		
1760	75		
1770	96		



https://plot.ly/~ktopham/22/number-of-books-in-the-top-10-subject-from-each-decade/

- 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**

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