Creating Databrarianship: Perils, pitfalls, and pratfalls of editing a research collection

Kristi Anne Thompson
Creating Databrarianship

Or, Herding Cats and Librarians:
Perils, pitfalls, and pratfalls of editing a research collection
My assistant
There's a hole in the middle of my book

Best-laid plans...
In the beginning... ACRL needed a book

- Which meant they needed someone to write or compile it for them
- First they contacted highly qualified librarians from major U.S. research institutions who sensibly said “no way” (but suggested other names)
- Which is how they ended up approaching Lynda Kellam (UNC Greensboro) and me
- We were initially asked to put together a proposal for them to consider, which needed to include outline of approach, topics covered, and a partial list of authors
Original concept had five sections

1. **Data Support Services for Researchers and Learners**
   - Data reference, data consulting / “wrangling”, studio model, embedded model...

2. **Data in the Classroom**
   - Quantitative literacy! Curriculum integration! Statistical literacy as part of information literacy!

3. **From Digital Maps to Data Miners: Data in the Disciplines**
   - Types of data: geospatial, qualitative, scientific, big

4. **Management and Preservation**
   - And documentation and citation and....

5. **Data Librarianship: Past, Present, and Future**
   - History, international perspectives, teaching the next generation, other...
Call for submissions got both interest and pushback

• Some chapters we had specific people in mind for, and we approached them directly. We also had vague “something about this” ideas we threw out to see what would happen, and left room for things we hadn’t thought of at all.

• Comment from librarian outside the data community: “oh, you're editing a book on research data management?”
  • Apparently assuming there was nothing else in data worth writing about

• From a data insider who read our prospectus: she “guessed” it was reasonable to include a whole section on data management
  • Implication being that data curation was some sort of passing fad

• Some people didn’t like the term “databrarianship”...

• But overall we were very happy with the response
Data in the Classroom

• Quantitative literacy is a topic dear to my heart...

• This section isn’t in the final book. Why?

• Centerpiece of the section was to be a discussion of incorporating quantitative literacy into the curriculum – and we had an author in mind:
  • A professor who taught that exact thing, had spoken on it at library conferences, and had expressed interest in collaboration between librarians and faculty

• He agreed with what appeared to be enthusiasm and we included his chapter in the proposal
Unfortunately, we didn’t have a backup in mind.
Final outline

1. Data Support Services for Researchers and Learners
   - Data reference, data consulting / “wrangling”, studio model, embedded model, quantitative literacy, some management services too

2. Data in the Disciplines
   - Geospatial, qualitative, scientific, big

3. Data Preservation and Access
   - Management and preservation and documentation and citation and....

4. Data: Past, Present, And Future
   - History, international perspectives, teaching the next generation
So what are we talking about anyway?

Anything that looks like a dataset to you.
Definitions: Data

• What is data? What are data?
  • We went with the plural and I don’t want to hear any more about it.

• Initial plan: cover in the (to be written much later) introduction; let chapter authors provide additional explanation as needed.
• After a few plaintive emails and a couple of chapter submissions with attempts (including one that was over 4 pages long) we realized that was not actually a good plan.
Some formal definitions

• “Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation.”
  OED

• "Reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.”
  ISO/IEC 23821: Information Technology Vocabulary
Google took a stab at it...

What qualifies as a dataset?
For purposes of inclusion, we take a broad view of what qualifies as a dataset:

- A table or a CSV file with some data
- A file in a proprietary format that contains data
- ... Anything that looks like a dataset to you

https://developers.google.com/search/docs/data-types/datasets
From the data community

From a debate on the topic on the IASSIST mailing list:

- “My historic definition of data is ‘information in numeric form.’ I guess that needs to be broadened to include non-numeric information…”
- “…a group of numbers to which you must apply some type of statistical or mathematical technique to derive meaningful information.”
- “Data can be almost anything.”
- “Data refer to vast quantities of information.”

*Data Basics, 2012, Diane Geraci, Chuck Humphrey & Jim Jacobs*

- “Designation of a concept with consistently determined extension.”
Let’s try to unpack that last one...

Conclusion

- **New Definition**
  - Datum := Designation of a concept with consistently determined extension
  - Answers the question – What *is* data?

- Dan Gillman
  - [Gillman.Daniel@BLS.Gov](mailto:Gillman.Daniel@BLS.Gov)

- Frank Farance
  - [Frank@Farance.Com](mailto:Frank@Farance.Com)

or maybe not.
Anything that looks like a dataset

• What kinds of data are in the book?
  • Includes geospatial (raster) data, qualitative data, etc. so numeric-only definitions are out.
  • We also decided to jettison “can be almost anything” as not useful.

• What do data librarians do with the data they work with?
  • Activities include: collecting, cataloguing, documenting, analyzing, interpreting, archiving, and converting between formats.
  • Datasets are entities (like books!) with creators

• Definition needed to specifically include types of data appearing in the book that could have these activities performed on them.
Our version

The term “data” is used in a variety of ways depending on field and use. A computer scientist might use the term to refer to the flow of zeros and ones... A statistician or survey researcher might think of a numeric dataset structured for use in a statistical package. In general, we will take an intermediate approach. Content on the web, video streams, survey responses, or engineering measurements constitute potential data for researchers, but the data we are concerned with here is the product of taking that raw informational input and assembling it into a structured form for analysis. In other words, data is a product of researchers as well as an input for research. Research data collections (or datasets) are generally in electronic form and are accompanied by or incorporate metadata, or documentation that describes the structure and content of the data.

Unless specified otherwise, data will be taken to mean electronic files of information that have been collected systematically, structured, and documented to serve as input for further research.
To summarize:

In brief, data are the raw materials for research, produced through any systematic collection, organization and documentation of information for the purpose of analysis.
The Histories of Data Librarianship

“A new role in academic libraries...”
Many authors had trouble figuring out where to start

• Some began *in media res*...
  • Data syntax files and fixed-with dictionaries require some introduction for most readers.
  • So do “wide format” vs. “long format”, “metadata extraction tools” and so on.
• Other authors took the approach that “there are no other chapters in this book and I need to start by explaining that libraries should be involved with data.”
  • Here comes data!
• Taking a historical approach was surprisingly common.
However, the people who took this approach couldn’t even agree on a start date

- “The data library, thought of as a collection of numeric and/or geospatial data resources for secondary use, emerged as early as the late 1950’s mainly in the United States, Canada, and the United Kingdom.”
- “In the early 1990’s data collections and services in Canadian academic libraries were in their infancy...”
- “In 1991, [article] stated data should not be ignored by the library...”
- “For decades, research universities and libraries have been developing services and support for working with administrative, research, and government data.”
- “Brief history of the data librarian
   Since 2003, many libraries have advocated for a job position that supports research data services...”
Also note that we already had two *entire chapters* devoted to the history of data librarianship, one looking at Canada (touching on North America generally) the other Europe (primarily the U.K.)
It’s a bit of a wonder that none of our authors ended up hating us by the time we were done…

Please find a different way to introduce your chapter.
Editorially speaking

Details, details...
Words, sentences, paragraphs

- Tense drifting
  - Past, present, future...
  - *Mostly* we wanted authors to pick a tense and voice and stick to it
  - Some had a little trouble...
  - One overdid it a little (using present tense to describe conducting a now-completed study, lessons learned, future directions)

- The “almost-right” words
  - Re-orient is not the same as re-evaluate
  - Accrete is not the same as acquire / obtain
  - Transpose is not the same as reshape

- Getting passive-aggressive on passive voice
  - Guilty.
Cover Story

- Why green blurry people?
- Our fault 😞
- Or maybe the graphics editor.
  - (she asked us...)
- We agreed: no 0’s and 1’s, no graphs, no other attempts to abstractly illustrate data.
- People? We disagreed...
  - So we ended up with blurry people obscured by a pattern.
Thank you!

Questions?