What Can You Do with a Digitized Book?

DH Methods for Analyzing Printed Materials

Anne Donlon, PhD
@annetiquate
What do you get when you digitize a book?
How do we use digitized books?
Annotation & Social Reading

Alan Liu, “From Reading to Social Computing,” Literary Studies in a Digital Age:

Similarly, blog and social-network platforms can be used to engage with literary works. For instance, one team of students in the 2008 version of my Literature+ course used the LiveJournal blogging platform to create profile pages for the characters in the General Prologue to The Canterbury Tales. They themed each profile page with images and language they thought appropriate to the character and then wrote comments “in character” on other characters’ pages or on a shared community page (e.g., the Miller commenting rudely on the Knight’s or Wife of Bath’s pages, much as Chaucer’s original Miller was wont to do).
What questions do we ask of digitized books at different scales?

What questions might we ask of one book?

What questions might we ask of every book published in the 1850s?

What questions might we ask of a million books?
Somewhere in there lies a manifesto for how digital humanities might reform certain academic orthodoxies that work against the hermeneutics of screwing around. Have we not already begun to call ourselves “a community of practice,” in preference to “a science, a method, a research, a pedagogy”?

But the real message of our technology is, as usual, something entirely unexpected—a writerly, anarchic text that is more useful than the readerly, institutional text. Useful and practical, not in spite of its anarchic nature, but as a natural consequence of the speed and scale that inhere in all anarchic systems. This is, if you like, the basis of the Screwmeneutical Imperative. **There are so many books. There is so little time.** Your ethical obligation is neither to read them all nor to pretend that you have read them all, but to understand each path through the vast archive as an important moment in the world’s duration—as an invitation to community, relationship, and play.

What do DH methods allow us to ask of digitized books?
TEI (Text Encoding Initiative)

Text Encoding and Libraries
The efforts of libraries to digitize their collections over the past several decades have been the root of librarians’ collaborative research support for digital humanities. In this work, text encoding became one of the initial and most prominent paths that libraries took to explore the connections of digital humanities to their services.

Harriet Green, “Facilitating Communities of Practice in Digital Humanities: Librarian Collaborations for Research and Training in Text Encoding”
TEI (Text Encoding Initiative)

- Began 1987
- Community of practice: TEI-L; TEI annual meeting; TEI Guidelines
- XML / markup language
- Structured data: tags & attributes define anything from a character name to a handwritten annotation to the title of the book.
- Used for both analysis and presentation (e.g., How many times does a female character speak to a male character? Format all chapter titles as header 1).
TEI, continued

From TEI by Example,
http://teibyexample.org/modules/TBED05v00.htm
Critical editions

- Willa Cather Archive
- Shelley-Godwin Archive
- Fluid Text Edition of Walden, Digital Thoreau
Counting words

Concordances, n-grams, word clouds, etc. with Voyant

- Emory University, Readux
  - Ladies First
  - A Short Account of the Rise and Progress of the African Methodist Episcopal...
  - An Autobiography the story of a colored evangelist...

Corpus analysis with AntConc, Programming Historian
Topic modeling

With some tinkering, MALLET generated a list of thirty topics comprised of twenty words each, which I then labeled with a descriptive title. Below is a quick sample of what the program “thinks” are some of the topics in the diary:

- **MIDWIFERY**: birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient
- **CHURCH**: meeting attended afternoon reverend worship foren nr famely performd vers attend public supper st service lecture discoarst administred supt
- **DEATH**: day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn
- **GARDENING**: gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds
- **SHOPPING**: lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower
- **ILLNESS**: unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach
Natural language processing

- **Natural Language Toolkit** (Python), **NLP** (R)
- Part-of-speech tagging
- Named entity recognition (identifying names of people and places)
- Sentiment analysis
- Rooted in discipline of linguistics
• Close Encounters
  - **Searching for Meaning** (searching variant word forms and word meanings)
  - **Parts of Speech** (analysing parts of speech (nouns, adjectives, verbs, etc.) of documents)
  - **Repeating Phrases** (analyzing repeating sequences of words)

• Distant Encounters
  - **Sentiment Analysis** (measuring opinion or mood of texts)
  - **Topic Modelling** (finding recurring groups of terms)
  - **Document Similarity** (measuring and visualizing distances between documents)

• Utility Examples
  - **Simple Sentiment Analysis** (measuring sentiment with a simple dictionary in the notebook)
  - **Complex Sentiment Analysis** (using research dictionaries to measure sentiment)
  - **Collocates** (identifying collocates for a target word)
  - **Concordances** (generating a concordance for a target word)
  - **Exploring a text with NLTK** (shows simple ways you can explore a text with NLTK.)

Table of Contents, *The Art of Literary Text Analysis* by Stéfan Sinclair & Geoffrey Rockwell
"Sentiment Analysis,"
The Art of Literary Text Analysis by Stéfan Sinclair & Geoffrey Rockwell

Then hate me when thou wilt; if ever, now; Now, while the world is bent my deeds to cross, Join with the spite of fortune, make me bow, And do not drop in for an after-loss: Ah! do not, when my heart hath 'scap'd this sorrow, Come in the rearward of a conquer'd woe; Give not a windy night a rainy morrow, To linger out a purpos'd overthrow. If thou wilt leave me, do not leave me last, When other petty griefs have done their spite, But in the onset come: so shall I taste At first the very worst of fortune's might; And other strains of woe, which now seem woe, Compar'd with loss of thee, will not seem so.
Stylometrics

- Genre detection:
  - HathiTrust
  - Ted Underwood
- Style of a particular author
Analyzing metadata

Dean Farrell, DPLA Metadata Explorer:  
http://www.deanfarr.com/metadata_explorer/#/dpla

Metadata for Women Writers Online:  
http://www.wwp.northeastern.edu/wwo/texts/metadata.index.html

Non-consumptive use of texts:  
https://analytics.hathitrust.org/datasets#genre
Mapping & spatial analysis

- Structured: tag place names in TEI
- Unstructured: detect named entities with algorithm, geoparser and/or gazetteer
- Z-Axis Tool
- NYPL, Navigating the Green Book
Network graphs

- Characters who speak to one another
- Characters who appear on stage together
- Citation networks
- Correspondence networks
- Network Analysis of Dramatic Texts
Beyond the Text

- **Book Traces**
- **IIIF** (International Image Interoperability Framework)
- Gifs and public domain image collections
  - British Library on Flickr
  - DPLA’s annual Gif it Up contest
- Image processing & analysis
How do librarians participate?

- Library of Congress Labs
  - “Recommendations for Enabling Digital Scholarship”
- NYPL Labs
- DPLA for Developers
- HathiTrust Research Center
How do librarians participate?

Collections as Data: Collections as Data Facets

- HathiTrust Research Center Extracted Features Dataset
  - “For every volume: metadata, including bibliographic metadata, word counts, and page counts.
  - “For every page in a volume: part-of-speech tagged tokens (words) and their counts. Metadata, including information about the page (number of lines, number of empty lines, counts of characters beginning and ending lines), and the language, which has been computationally determined.”
Access, impact

“If other institutions are anything like mine, we're killer at counting digitization throughput, and measuring analytics around the performance of the materials we put online in our Digital Collections platform, or the HathiTrust Digital Library, or DPLA, etc. But those are just shadows on the wall compared to what we're really after, which is meaningful engagement with and use of the materials.”

Josh Hadro, “Manufacturing Impact: Why We Digitize,” NYPL.
Questions?

Resources:

- DH + Lib
- Digital Library Federation, especially #DLFteach working group
- DH Slack
- TAPoR (Text Analysis Portal for Research)
- Dirt Directory
- BH and DH
- Digital Humanists on Humanities Commons
- Journal of Interactive Technology & Pedagogy