Extracting Metadata from Digital Records Using Computational Methods

Kate Barbera
kbarbera@andrew.cmu.edu
@brightarchives

Ann Marie Mesco
mesco@andrew.cmu.edu
@amarieannm

Investigation

This project aims to develop an automated, scalable workflow for extracting item-level metadata from digital records using tools and technologies employed by the community (archivists, digital humanists, etc.). In 2016, the Carnegie Mellon University Archives began research in this area as part of a large repository migration. Our digital collections have 3 million+ pages of items, and a repository-wide assessment found the metadata to be widely inaccurate and inconsistent. Due to the size of the digital collections, traditional refinement methods proved impractical.

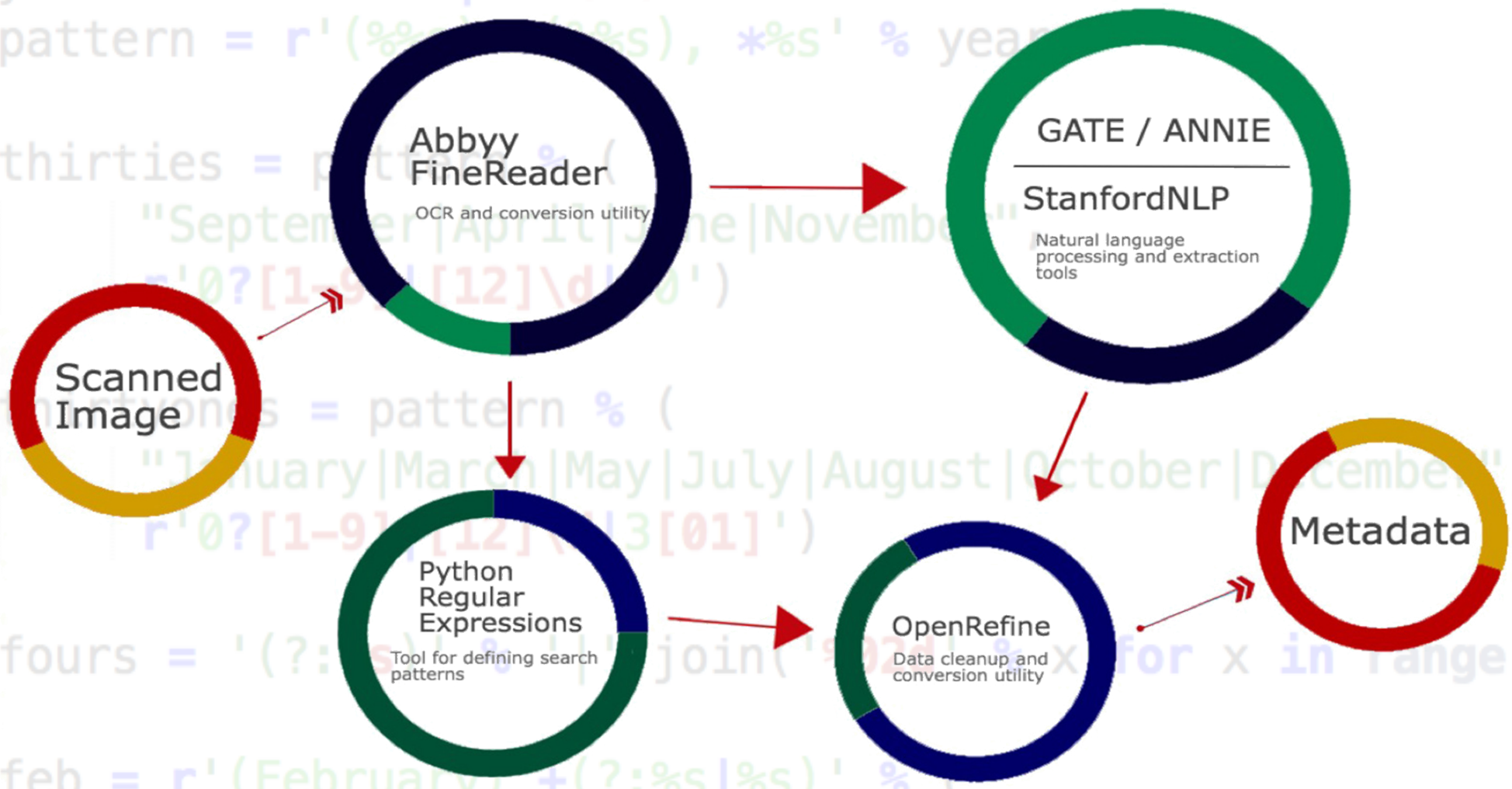
Can we efficiently integrate this workflow into our current practices?
How do we scale from pilot to program?

Case Study

Test workflow on the William W. Cooper Collection (2,884 items) by:

- Evaluating existing OCR files and cleaning resulting text when necessary.
- "Categorizing" records based on genre, form, and other characteristics (e.g. - correspondence).
- Using scripting tools (Python, RegEx) to highlight and extract key metadata values (title, date, creator, etc.).
- Employing Natural Language Processing (NLP) tools to identify potential subject headings and/or key words.
- Using OpenRefine, DataWrangler, etc. to clean and normalize resulting metadata values.
- Comparing research workflow with existing local practices.

Workflow



Challenges

- Excluding any scientific and mathematical equations, the OCR files for the case study are roughly 70% accurate.
- The case study uncovered more than 14 genre or form types. The largest—research reports—contains 1041 items, while the smallest—drafts—contains 9.
- If there are less than 100 records in a "category" (i.e. - genre/form), traditional methods are more practical and efficient (e.g. - 10 minutes per record across 50 items equals roughly 8 hours of work).
- This method is most useful for metadata values that cannot be generated or modified in large batches (title and date).
- Unlikely to achieve majority accuracy (more than 50%) using this method without further refinement and normalization.
- Fixed resources and limited opportunities for training.

Future Research & Goals

- Test this workflow on additional properties such as creator.
- Scale this workflow for larger collections (300,000+ items in the H. John Heinz Collection).
- Explore topic modelling as a method for extracting additional subject headings and/or key words.
- Investigate improved OCR technologies for scientific and mathematical formulas.
- Explore crowdsourcing solutions for normalizing and enhancing resulting metadata values.
- Share all resulting research and tools via GitHub repository.
- Continue to pursue accessible, practical solutions that we can share with the broader community.

https://github.com/cmuarchives/metadata.git
http://digitalcollections.library.cmu.edu

Special thanks to Dr. Jessica Ottis

Carnegie Mellon University