Experiment: Exploring Computational Description

Creating standardized catalog records from ebooks and other digital materials

Partner Service Unit: LCSG | Use case: LC20230002

Initiated: 12/2021 | **Begun**: 09/2022 | **Final deliverable**: 04/2023

AI Techniques Used: Token classification, text classification, text extraction, human in the loop workflow, analysis and classification

Overview

The Exploring Computational Description (ECD) experiment considered the use of AI tools to help catalog ebooks. The quality and accuracy of algorithmically derived metadata from full-text ebooks were evaluated against the existing records created manually by Library catalogers using standard workflows. The viability and utility of five current machine learning (ML) models were also evaluated during the experiment.

Overall, the experiment supported three major conclusions that will inform future iterations and related experiments: 1.) certain fields -- like identifiers (LCCN, ISBN), author, and title – readily lend themselves to useful extraction, while others like subject headings, genre, and dates proved difficult to reliably distill; 2.) the success of this experiment (and others like it) is highly dependent on the quality and robustness of the training data, particularly for non-generalist tasks like extracting library metadata; and 3.) no current AI tools are ready for production at this point, but this is the right time to increase internal capacity to meet the tools as they improve.

Context

The library acquires hundreds of new ebooks daily. While modern cataloging workflows are the product of decades of refinement and best practices, the explosion of digital acquisitions presents an important opportunity to evaluate how AI tools might support cataloging evolution. Of particular interest is how basic data can be extracted from texts algorithmically, leaving valuable cataloger attention for more complex and subjective matters deserving of expert handling.

ECD was sponsored by the Principal Deputy of the Librarian of Congress and co-led by Caroline Saccucci, Chief of the U. S. Programs, Law & literature Division in ABA/DPS/LCSG and Abigail Potter, Sr. Innovation Specialist in LC Labs/DID/DSD/OCIO. It was the first task order utilizing the Digital Innovation IDIQ which provides a structured way to experiment and learn about the effectiveness of AI technologies with Library data and use cases.

Hypothesis

The original research question for the experiment:

> What are examples, benefits, risks, costs and quality benchmarks of automated methods for creating workflows to generate cataloging metadata for large sets of Library of Congress digital materials? And, what technologies and workflow models are most promising to support metadata creation and assist with cataloging workflows? What similar activities are being employed by other organizations?

Process

The experiment asked the contractor to test and report on at least five machine learning models to generate cataloging data using the textual and/or visual elements of ebooks in epub, PDF, or other digital

formats as an input source. Approximately 23,000 cataloged ebooks were provided for model training and to serve as ground truth, i.e. the basis for validation and evaluation. A variety of models were trained, tested, and tuned.

Performance scores for individual fields were based on how well the models predicted a result that matched the existing catalog records, as illustrated in Figure 1. The models were extremely good at identifying specific text strings as identifiers, e.g. 010 LCCN and 020 ISBNs, as well as the names of authors (100) and titles (245), with an accuracy rate of over 95%. The models were much less successful in predicting genre terms, with an accuracy rate of less than 30%.

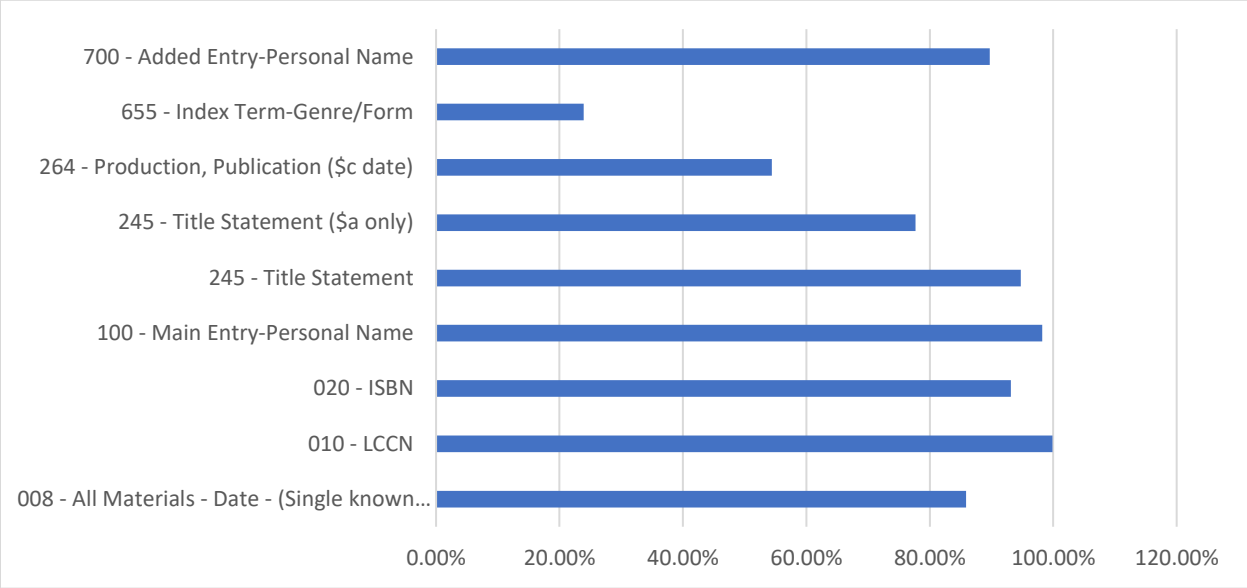

Figure 1: Combined accuracy rates for selected MARC fields

The contractors also developed experimental interfaces that presented AI-generated suggestions for subject headings and name authorities that could augment cataloging workflows at the Library and enhance access to Library collections.

Risks

The main risks associated with automated description tasks are:

- **Error:** there's a reputational risk for the Library if books are miscataloged and unable to be found or associated with the wrong author/title/identifier.
- **Bias:** algorithmically produced data may display systematic bias in multiple ways, reinforcing bias latent in training data and current systems.
- **Instability**: the AI landscape is subject to frequent upheaval with new frameworks and models are coming online all the time; to date these tools have immature tooling and affordances for integrating into domain-specific workflows.

Other Conclusions and Lessons Learned

- Co-leading this experiment with ABA has been crucial to fully exploring the use case and the data.

- Documentation requirements for vendors are proving exceptionally valuable for increasing staff understanding and ability to plan for future work.
- The experiment raises valuable questions about quality standards.

Value to Congress

- Extensive engagement with state-of-the-art industry technologies in a controlled setting which gives the Library needed experience developing training sets and means of evaluation.
- A better understanding of where to focus Library staff expertise and piloting of “humans-in-the-loop” processes for supporting and extending expert workflows.

What’s Next

- A second iteration experiment will refine quality standards and assessment methods for applying machine learning methods to the creation of catalog records.
- Further exploration of other areas for machine-assisted metadata extraction (see also: Classifying data for congress.gov using Machine Learning, and Historic Copyright Records).
- Collaborating broadly to refine AI specific planning, risk rubrics, and templates with other LAM colleagues.
- Supporting service units and Library staff who have expressed interest in the topic, experiments, or domain profiles.