Caroline Saccucci and Abigail Potter

16 Assessing Machine Learning for Cataloging at the Library of Congress

Abstract: Born-digital content is being produced by all and collected by libraries at an increasing rate. Workflows for cataloging this material continue to evolve and have the potential to be augmented by automated technologies. To explore the potential of machine-generated cataloging methods, the Library of Congress is undertaking phased research and experimentation through a specific project: *Exploring Computational Description.* First, using thousands of cataloged ebooks, five different machine learning models are being trained, tested, and documented. The most promising approaches will be applied to uncatalogued ebooks and evaluated. Building on the first phases of experimentation, additional potential automated workflow improvements will be tested. The process of evaluation will result in foundational quality benchmarks for automated methods along with detailed benefits, risks, and costs. Outcomes of the experiment will be used to inform the future development of born-digital cataloging workflows at the Library of Congress and potentially elsewhere. The experiment utilizes a set of tools developed by the LC Labs team to plan, document, analyze, prioritize, and assess AI technologies. The chapter includes a review of the AI planning tools being used and reports on the experiment in progress.

Keywords: Machine learning models; Data processing documentation; Cataloguing workflows

Introduction

In August 2022, the Library of Congress (LoC) of the United States initiated an experimental project, *Exploring Computational Description*, with the help of a third-party vendor, Digirati, to investigate machine learning (ML) processes to create or enhance bibliographic records for born-digital content. The experiment, jointly coordinated by staff in the Library's Office of the Chief Information Officer and the Acquisitions and Bibliographic Access Directorate, provides a mechanism for evaluating five potential models for using ML methods in metadata description and two cataloging workflows to assist catalogers in describing digital content. This chapter describes the background and early details of the experiment.

https://doi.org/10.1515/9783111336435-017

Artificial Intelligence and Machine Learning: The Next Wave of Technological Change

Machine learning (ML), a subfield of artificial intelligence (AI), was defined in the 1950s as "the field of study that gives computers the ability to learn without explicitly being programmed." ML processes are trained to recognize patterns, predict patterns, and make suggestions about what actions to take (Brown 2021). Machine learning is dependent on algorithms or models that are trained on data. Very simply, training data are tagged or marked up according to the desired pattern. The models are trained, tested, tweaked, and retrained on training and test data, then applied to larger volumes of untagged target data, and asked to recognize similar patterns.

ML technologies have been in use in libraries, archives, and museums (LAMs), primarily in the form of Optical Character Recognition (OCR), for decades. Providing machine-readable historic texts using OCR techniques to capture the content of books and manuscripts has enabled the digital humanities to grow into a recognized world-wide field of study. Applying data science techniques to large corpuses of digital text has spurred new kinds of analysis and research. OCR has also enabled powerful search and discovery tools to connect digital collections to a wide variety of users. Advances in the capabilities and availability of ML tools that could be applied to a larger scope of LAM content beyond text, such as images, audio-visual, born-digital manuscript collections, web archives, and recorded sound, could have similar broad impacts.

Specific Challenges for Libraries, Archives and Museums

The LAM community and LC Labs, a digital strategy team within the Library's Office of the Chief Information Officer, have been intentionally collaborating, experimenting, and sharing results of small-scale AI projects since 2019 (LoC LC Labs n.d.). There is a universal challenge in moving any of the small-scale experiments to operationalized technologies. Most institutions do not currently have the technical expertise or literacies to develop and test custom AI tools, and they must rely on vendor-provided or commercial solutions. Typically, commercial, or free-use options are not open source, and specifics around the nature of the models and training data that are utilized are not shared. Commercial tools and data may also only be available when used on a commercial cloud platform.

In tests by the LAM community, many of the widely available commercial tools do not perform well with historic, digitized, or formatted materials. As noted in a report on the LC Labs experiment using the Amazon transcribe tool and a Speech to Text viewer, the transcripts based on "regional and older styles of speech" were inaccurate because the particular style of speech was likely not present in the training sets (Adams and Kim 2020).

> These issues are inherent to some collections; even listening and understanding what is being said can be very difficult. Additionally, many items had gone through previous necessary physical carrier migrations before being digitizing, resulting in . . . poor "signal-to-noise" ratios, with artifacts like crackling. The accuracy of contemporaneous born-digital audio, in contrast, was high (Adams and Kiim 2020).

The report David Smith and Ryan Cordell released in 2018: *A Research Agenda for Historical Multilingual Optical Character Recognition*, called out this same problem with text-based OCR stating "While large-scale scanning projects have generally used off-the-shelf OCR products, several researchers on a smaller scale have found that domain-specific training and modeling provide significant gains in accuracy" (Smith and Cordell 2018, 11).

LAMs are sources for authoritative and trusted information and they act for the public benefit. The collections they steward are complex and contain a wide variety of physical and analog formats with restricted, private, or sensitive content. Highly trained staff with deep knowledge and expertise have always been, and will continue to be, the bridge between collections, services, and the public. Automated technologies could disrupt the operational principles of LAMs. For example, errors in description about sensitive content that show up on a Library of Congress MARC record could have greater repercussions for users, staff, and the organization than that same error showing up on a search result page of a commercial search engine. The machine-readable cataloging record format (MARC) has driven automated cataloguing developments since the 1960s and continues to do so (Library of Congress 2023). However, a justified desire to implement AI responsibly coupled with a lack of direct experience or practical guidance on implementing AI in LAMs might slow down adoption.

To move beyond small-scale experimentation in AI and take an active role in how this influential technology will shape the field, LAMs must develop shared quality standards, governance structures, and clear requirements for how AI tools need to perform to support the content and principles that are inherent to LAMs. The values and aspirations which shaped the adoption of digitization and digital preservation technologies must be reflected in AI implementation within LAMS. The new environment requires the development of targeted tools to understand the specific risks, benefits, and mitigation approaches for implementing AI at a human-scale and perhaps at a slower pace.

Developing Human-centered and Domain-specific Applications of Artificial Intelligence Responsibly

The LOC is not alone in considering ways to implement AI technologies responsibly. The community of technical and subject matter experts brought together by the Artificial Intelligence for Libraries, Archives and Museums (AI4LAMs) group is a network of peers from similar organizations who are sharing the lessons learned as they experiment with AI (AI4LAM n.d.). AI4LAM sponsors events, comprises working groups and chapters, and produces an Awesome List of AI Resources. The Office of the White House in the United States Government has made a call for articulating the concerns and rights of humans in AI systems (White House 2022). Its blueprint for an AI Bill of Rights embraces five principles: safe and effective systems; algorithmic discrimination protections; data privacy; notice and explanation; and human alternatives, consideration, and fallback.

Various actions by organizations add credibility to the emerging best practices around creating documentation for AI data. The Data Nutrition Project "seeks to create tools and practices that encourage responsible AI development, partners across disciplines to drive broader change and builds inclusion and equity into our work" (n.d.). It has produced a dataset nutrition label to ensure transparency, seeking to replicate the nutrition labels on food. A group within the ML community has proposed the use of a datasheet for each dataset, mirroring the datasheets in the electronics industry. The dataset datasheet would include composition, collection process, and recommended uses (Gebru et al 2021). Further work on ensuring responsible use of AI has been undertaken at OCLC. A research agenda proposes seven areas for investigation including commitment to responsible operations, sharing of methods and data, machine-actionable collections, workforce development, and interprofessional and interdisciplinary cooperation (Padilla 2019). The activities of the various groups directly influence and inspire the development of the LC Labs AI planning and assessment tools.

Artificial Intelligence Planning and Assessment Tools

LC Labs has prepared planning and assessment tools to be used on their various projects. The goals of the tools are to gain specificity, establish baseline performance metrics, and build in pauses to assess project alignment with stated LAM principles and goals. Another important aspect of the tools is the act of gathering a diverse set of stakeholders with internal and external perspectives, including those groups who have the potential to be impacted by an AI system, to collaborate in the planning process. The tools are being used for the first time with the *Exploring Computational Description* experiment, and they will continue to be refined. Two of the tools are described in detail below. The two tools are the Organizational Profile and the Data Processing Plan.

Organizational Profile

The National Institute for Standards and Technology (NIST) AI Risk Management Framework (NIST 2023) includes an examination of the risks of AI implementations. These include reliability, validity, security, resilience, accountability, and transparency, explainability, interpretability, and fairness. The framework proposes core functions: govern, map, measure and manage, and use-case profiles. Building an organizational profile or functional profile to map and define potential uses of AI in an organization is a recommendation in the NIST framework. This step helps to define the specific AI tasks and methods in the context of an organization or its users.

For example, in an initial organizational profile for LC Labs, four functional areas emerged:

- Enabling discovery at scale
- Enabling research use
- Enhancing collections processing and data management for internal workflows and business cases, and
- Augmenting user services

The first, *Enabling Discovery at Scale,* relates to generating metadata for items, papers, articles, paragraphs, or objects to enhance search and discovery with example tasks of processing digitized collections with OCR, speech to text transcription, and named entity linking. The next *Enabling Research Use* emphasizes making data and guides available for researchers and other users to analyze and includes processes to create and process datasets or research corpora for use by external users. Users may request the creation of datasets, so they can run AI or ML techniques like natural language processing (NLP), text mining, or sentiment analysis. A library may also run the processes to answer specific researcher requests. The next area in the initial profile *Enhancing collections processing and data management for internal workflows and business cases* concerns the support of local content management, reporting, and analysis. A subcategory might be the management, processing, and preservation of born-digital collections, such as web archives, email archives, and other born-digital manuscript materials and include automatic classification, segmenting documents, creating automated workflows, and humans-in-the-loop (HITL) processing which involves combining human review with machine learning in a workflow. The final functional area in the initial mapping is *Augmenting User Services,* which includes tools for public-facing services like recommending systems, chatbots, and voice searching.

By mapping out and organizing specific tasks, it is possible to gain insight into the risks and benefits of a functional area that may be similar to or distinct from others. For example, in the area of collections processing and data management, the users of the system are internal staff, and feedback and input from staff are essential in designing systems. Additionally, high visibility tasks in collection processing areas are designed for different levels of staff oversight and HITL workflows. Mapping an organizational profile helps in prioritizing where to focus effort. Hypothetically, LC Labs has undertaken foundational experimentation in one of the functional areas, *Augmenting User Services*. If a broad base of experimentation were the goal, testing out technologies and gathering baseline information about voice search or chatbots could be a next step. The *Exploring Computational Description* experiment fits into the management and processing of born-digital collections sub-category. The users are catalogers and digital collection managers who constitute the people who will review deliverables from the project and integrate any findings into future planning.

Data Processing Plan

The Data Processing Plan (DPP) is a template that vendors, partners, or staff can complete to document data transformations, specifically transformations using AI or ML technologies. It brings together emerging AI documentation standards like Google's model cards (GoogleCloud n.d.) and data coversheets (Gebru et al. 2021) and includes LAM-specific sections for documenting data provenance, potential gaps in data, and potential risks to people, communities, and organizations. Risk management is a required deliverable for experimental data processing in Library of Congress contracts. An initial plan is required before LoC data are processed to outline the intent of the processing, the preparatory and processing steps, the descriptions of the data used in the experiment, and the models with expected performance information. At the end of an experiment, a final DPP is required to document the actual performance and delivered data.

At the time of writing, the initial DPPs have been delivered for the *Exploring Computational Description* experiment. They provide an incredible amount of detail and specificity about the ML models and data that are being tested. When the series of experiments is completed, the final DPPs will provide foundational information about performance requirements and expectations for specific models and how each performs. Compiled over time on a variety of models and LAM data formats, DPPs and other similar documentation can contribute to the development of shared quality standards for AI/ML processing.

Table 16.1 shows an excerpt of the DPP for the Annif model, one of the five models being tested and one of the initial DPPs submitted for the *Exploring Computational Description* experiment. Annif is a tool for automated subject indexing developed at the National Library of Finland.

Table 16.1: Data processing plan for Annif

1) Please describe the purpose of this dataset with relation to the ML/AI workflow. Explicitly address if it is being used as training, validation, or test data.

Where possible, we will use cross-evaluation when training models on LoC data in order to avoid introducing selection bias or overfitting the model to the training set. If this is not possible, the dataset will be explicitly split into training, validation, and test data without cross-evaluation. The split will be random, and follow a standard 80/10/10 split. We would expect the training, validation, and test data to comprise examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset. However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation, or test.

b) For training data:

1. if the model is pre-trained, describe the data on which it was trained
2. if the model will be fine-tuned, outline the data involved in this process
3. if the model is being trained from scratch, outline the plan for creating training data.

We would expect to:

1. *Train the model(s) based on the training subset of the LoC ebook dataset*
2. *Testing the LoC-trained models on a test subset of the LoC ebook dataset.*
3. *Produce scores/metrics for each record, and for the collection in aggregate for each testing cycle.*

Each of the training and fine-tuning steps will use the text from the books and the MarcXML records.

c) If creating training data or validating training data using volunteers or paid participants (e.g. via crowdsourcing), please describe the workflow and incentive structure.

This experiment does not include data generated or collected from volunteers or paid participants.

d) Document any known gaps in the dataset, such as missing instances or forms of representation. Address possible sources of vias in the dataset resulting from these discrepancies.
1. Describe any steps taken to remediate or address gaps or bias in a dataset used in the ML/AI processing or in the experiment overall.

For this experiment, the goal is to test, in a time-limited period, the success of these models in matching existing human catalogers at generating bibliographic metadata from ebooks. The type of task being carried out in this experiment is less likely to surface bias, as we are primarily looking for existing text in an existing record, and will be fine-tuning models based on existing catalog records.

To the extent that any biases show up in the data outputs, these will be reflected in lower scores (where the bias leads to misclassification).

Details of the *Exploring Computational Description* Experiment

Exploring Computational Description is a year-long experiment to test multiple ML models in their ability to generate MARC record catalog metadata from the contents of digital materials, specifically ebooks. The technical work is being done by the firm Digirati. It is the first in a series of ML experiments that are gathering baseline performance and quality data for generating priority catalog metadata. The research questions for the initial experiment are:

- 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?
- What technologies and workflow models are most promising to support metadata creation and assist with cataloging workflows? and
- What similar activities are being employed by other organizations?

Table 16.2 provides an overview of the experiment’s scope of research, targets for quality review and key deliverables.

Table 16.2: *Exploring Computational Description* Experiment

Scope of research	Targets for LoC quality review	Key deliverables
Test five ML models or methods to detect or generate full level bibliographic records. All models are open source	Expected generated fields: – titles – author names – unique identifier – date of issuance – date of creation – genre/form, and – subject terms	– Data Processing Plans – Performance reports + data – All data utilized or generated in the experiment
Test two additional ML techniques to augment cataloging workflows	– Subject classification workflow – Proper name disambiguation workflow	– Data Processing Plans – Rough prototype – All data utilized or generated in the experiment
Test most promising models on uncatalogued ebooks	– Review for potential use in LOC systems – Use for further tests	– Delivery of MARC21 and BIBFRAME metadata
What ML or other automated processes are similar organizations using?		– Findings and recommendations report

Data Involved in the Experiment

LoC has delivered data to the contractor to train and test the models. The training data consisted of a total of 23,130 items and their existing catalog records and included 13,802 Cataloging in Publication (CIP) titles, 5,835 open access ebooks, 403 edeposit ebooks, and 3,750 legal reports containing a mix of digitized and born-digital content. The five different models being tested are being trained on these data. The models yielding the most results will be used to generate MARC records for approximately 50,000 uncatalogued ebooks. The LoC will review and test the automatically generated records for potential use in its systems; however, it is more likely that the data will primarily be used for further experimentation.

Machine Learning Processes to Be Tested

Five ML models are being trained and tested in their ability to create high-quality MARC records. The description of the model capability is provided by the vendor in the initial DPPs. A mix of text extraction and visual analysis approaches will be tested. Each model, with some in combination, will be tested in how well each can generate the key MARC record fields. The five models are:

1. GROBID (GeneRation Of Bibliographic Data) is a machine learning library for extracting, parsing, and re-structuring raw documents such as PDF into structured extensible markup language (XML)/text encoding initiative (TEI) encoded documents with a particular focus on technical and scientific publications. The extraction includes bibliographical information, for example title, abstract, authors, affiliations, and keywords, along with the text and document structure. GROBID will be used to provide initial benchmarking with an off-the-shelf tool, and the model then trained to be more tailored to LoC data. GROBID is also useful for generating XML files which can be used for text input for subsequent experiments.
2. Annif: automated subject indexing toolkit, a tool from the National Library of Finland which is designed for automated subject cataloging. Annif provides access to multiple ML backends facilitating trials of different ML models and approaches, including term frequency - inverse document frequency (TF-IDF) and multi-modal language model (MLLM), and benchmarking a wide range of approaches to subject and genre cataloging.
3. Spacy. Spacy is an industry standard NLP library in Python, with extensive abilities to be trained and customized with additional pipeline steps for LoC catalog metadata, and can be used for the full range of metadata for the experiment, including subjects, genres, and bibliographic metadata.
4. Bidirectional Encoder Representations from Transformers (BERT) testing and training a wide range of BERT-derived LLMs including BERT, RoBERTa, and distilBERT, and transformer-based approaches for token classification, identifying words or phrases in text like titles, authors, and dates, and for text classification which classifies a body of text like assigning subjects and genres.
5. NLP with Layout features. The use of this approach would supplement either the fourth or the fifth model, depending on the outputs of the earlier experiments, with layout data such as page position, text size, text location, page number, and recto/verso, to identify whether visual information can add additional weighting to the NLP models and to further refine data extraction for titles, authors, and other fields that have distinct positions, or formatting within the document.

DPPs accompany each model to be tested and the results will provide baseline information on which models perform well with LoC data and which models do not. LoC catalogers and digital collection managers will review the data and use the insights for further experimentation and to inform technical and performance requirements for potential future systems.

Assisting Cataloging Workflows

Based on the performance of the models, the LoC has asked the vendor to select and test an additional two ML methods to assist catalogers in workflows for digital content. After a series of stakeholder and user workshops, the following ML workflows were selected:

- Text Classification which would generate subject and genre data labels from the text and supplement the outcome with text summaries and taxonomies for subject classification, and
- Token Classification which would identify specific entities within the text and generate data and keywords, including coinciding entities, and use data from the Program for Cooperative Cataloging (PCC)
- Name Authority File and elsewhere to disambiguate named authorities.

The tests will result in a basic prototype which catalogers will use to select ML-generated terms. LoC catalogers and digital collection managers will review the output and provide feedback on the information provided in the prototype. Insights will inform future experiments and planning efforts for future development of born-digital cataloging workflows.

Next Steps

The next steps in the experiment are to review the performance reports and data from the tests, define quality review criteria and plans, pause for assessment and alignment, plan for future task orders, and share outputs from the experiment widely.

Conclusion

Experimenting with machine learning models for bibliographic description enables the Library of Congress to make strategic resource decisions for cataloging large quantities of digital content. Contracting with an external vendor for experimental work provides the LoC with experts in machine learning to test out various possibilities without taking staff resources away from production cataloging. As the current experiment progresses and with follow-on experiments waiting to be implemented, the Library of Congress will continue to explore computational description while assessing machine learning to provide discovery and access to even more of the Library's rich digital resources.

References

Adams, Chris, and Julia Kim. 2020. "Experimenting with Speech to Text and Collections at the Library." *Library of Congress Blogs. The Signal Blog: Digital Happenings at the Library of Congress*. Posted by Leah Weinryb-Grohsgal. June 18, 2020. https://blogs.loc.gov/thesignal/2020/06/experimenting-with-speech-to-text-and-collections-at-the-library/.

Artificial Intelligence for Libraries, Archives and Museums (AI4LAM). n.d. http://ai4lam.org/.

Brown, Sara. 2021. "Machine Learning, Explained." MIT Sloan School of Management. April 21, 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained#:~:text=

The Data Nutrition Project. n.d. https://datanutrition.org/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughn, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." V8. arXiv:1803.09010. Top of Form https://arxiv.org/abs/1803.09010.

GoogleCloud. n.d. "Model Cards: The Value of a Shared Understanding of AI Models." https://modelcards.withgoogle.com/about.

Library of Congress (LoC) LC Labs. n.d. "AI at LC." https://labs.loc.gov/work/experiments/machine-learning/.

Library of Congress (LoC) Network Development and MARC Standards Office. 2023. "MARC Format for Bibliographic Data." 1999 Edition Update No. 1 (October 2000) through Update No. 37 (December 2023). https://www.loc.gov/marc/bibliographic .

National Institute of Standards and Technology (U.S.). 2023. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)". Washington, DC: US Department of Commerce NIST. January 31, 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Padilla, Thomas. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97.

Smith, David A., and Ryan Cordell. 2018. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Boston MA: Northeastern University NULab for Texts, Maps & Networks. With the support of the Andrew W. Mellon Foundation. https://repository.library.northeastern.edu/downloads/neu:m043p093w?datastream_id=content.

The White House Office of Science and Technology Policy (OSTP). 2022. "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." October 2022. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.