Feasible, Adaptable, and Shared:

A Call for A Community Framework for Implementing ML and AI

|  |  |  |  |
| --- | --- | --- | --- |
| **Abigail Potter** | **Meghan Ferriter, PhD** | **Eileen Jakeway Manchester** | **Jaime Mears** |
| *Library of Congress**USA**abpo@loc.gov**ORCID* | *Library of Congress**USA**mefe@loc.gov**ORCID* | *Library of Congress**USA**ejakeway@loc.gov**ORCID* | *Library of Congress**USA**jame@loc.gov**ORCID* |
|  |  |  |  |

**Abstract – Through open research, experimentation and convenings with LAM sector peers and colleagues, a foundational need has emerged for a broadly shared and evidenced set of guidelines for implementing ML and AI technologies that centers the long-term stewardship and ethical responsibilities of cultural heritage organizations. Inspired by community guidelines that rationalize complex information into an understandable framework like the NDSA Level of Digital Preservation and the Data Nutrition Project, LC Labs is proposing a step toward collaboratively generating a LAM-specific framework for understanding and implementing ML and AI technologies.**

**Keywords – Community guidelines, machine learning, artificial intelligence, transparency, experimentation**

**Conference Topics – Community, Resilience**

# Introduction

Another wave of technical change is at the door of many libraries, archives and museums (LAMs). The promise and claims of artificial intelligence (AI) systems to transform organizations and solve entrenched challenges with data-driven results and solutions are enticing. Especially when users are expecting consistent and sophisticated search and discovery systems across all formats and content types. In addition to the shared challenge of user expectations, cultural heritage and research organizations have limited budgets, technical staff and expertise in implementing AI-driven services. As a result, formal and informal networks are forming to develop and share strategies and practices for dealing with this latest wave of transformation.

Through open research, experimentation and convenings with LAM sector peers and colleagues, a foundational need has emerged for a broadly shared and evidenced set of guidelines for implementing ML and AI technologies that centers the long-term stewardship and ethical responsibilities of cultural heritage organizations. Inspired by community guidelines that rationalize complex information into an understandable framework like the NDSA Levels of Digital Preservation [1] and the Data Nutrition Project [2], LC Labs is proposing a step toward collaboratively generating a LAM-specific framework for understanding and implementing ML and AI technologies.

# Resilience through community

Despite outstanding efforts to digitize and preserve historical materials, the information they hold remains difficult to use computationally, fragile to sustain, and unwieldy for systems that serve modern user needs. AI systems hold promise for solving our technical and data challenges. However, digital library, archive and museum collections generally need to be transformed to be used by data-centric technologies and tools like machine learning (ML) and artificial intelligence (AI). And these systems are generally sold by vendors. In any AI or ML process there is potential for distortion or loss of context at each stage of transformation; this risk is exacerbated when proprietary algorithms are used. Understanding the potential for generating positive impacts for the users of LAM collections verses the potential for harm in using untested technologies has spurred the Library of Congress Labs team (LC Labs) to sponsor public experiments in machine learning to gather evidence about benefits and risks before making large scale investments in implementing what is often proposed by vendors as a soup-to-nuts AI solution.

With each technical advance, be it creating machine-readable bibliographic records, digitizing collections, making content available online, navigating digital publishing and social media, and managing and preserving digital collections, the LAM community has developed shared tools, practices and standards to respond to a changing technical landscape. These tools and standards, like the MARC record standard, FADGI digitization guidelines, the WARC format for preserving web archives, and various file format identifiers were developed through trial, error and committee agreements to benefit users, improve institutional practices and give guidance to staff who often have to train themselves on the latest technologies and advancements.

# Community AI Practice

The LAM community has seen immense benefit from reports demonstrating the imperative of ethical adoption and research agendas to inform use and proliferation of AI in cultural heritage [3] [4] [5] [6]. In addition to these resources, events and workshops have brought together practitioners and leadership to highlight the practical challenges in this problem space [6] [7] [8]. Ongoing communities of practice continue to share the outcomes of their regularly convening, including the AI4LAM community. Additionally, there are vibrant existing and emerging disciplinary collaboratives, which present opportunities for the LAM community to engage more deeply around how ML and AI can perpetuate legacies of silence, harm, and structures of power. These activities offer the potential to synthesize practices and facilitate knowledge exchange and evaluation.

Concurrent to these community initiatives and knowledge sharing activities, LC Labs sponsored research and experimentation has generated evidence, surfaced complexities in applying methods, and produced recommendations shared widely to benefit the community. Through experimentation, research, collaboration, and reflection, LC Labs works to realize the Library’s vision that “all Americans are connected to the Library of Congress” by enabling the Library’s Digital Strategy [11]. While pursuing this line of experimentation and convening practitioners, LC Labs staff have encountered challenges shared by a wider community. Before discussing LC Labs experimentation toward a community framework for ML and AI, we will briefly summarize some of those challenges.

# Challenges Toward a Framework

Despite these promising activities and the shared needs surfaced from these community activities and events, challenges remain. The landscape of available and effective methods is rapidly evolving, as are organizations as they test and even expand capacity to adopt and implement these methods. A shared framework would likely address a range of challenges, including these challenges in taking a first set of steps into this practice.

Distortion of and the loss of context in the production of digital collections are issues [12] that go back to collection acquisition, selection, description, digitization, management, online availability and then mass digitization. One of the key challenges of the transformations done by ML and AI technologies is the lack of transparency in decision-making at the human and systems level. Interpreting viability and nuance within the results of ML applications requires human expertise, as well as clear articulation of each step in a project’s lifecycle; to include decisions of inclusion, exclusion, availability, and source and training data dimensions.

Experimentation and iteration should be essential to adopting approaches and a framework and its support for implementation. However, we acknowledge that creating space and leadership buy-in for experiments and pilots—and even prototyping—is more complicated in practice than on paper. At this time, it is particularly difficult to convey–in advance of undertaking initiatives–consistent predictions about resource requirements, risk, complexity, and user and organizational needs; precisely because these types of information are gathered through the process of undertaking this work. Following the completion of an ML or AI project, it can be a challenge to immediately assess impact and define coherent next steps in advance of broader evidence.

Moving beyond project level implementation to more systematic exploration remains a specific challenge in fields in which resources have not yet been allocated for wider programmatic implementation. Additionally, as frequently shared outcomes and related effective practice tend to represent discrete projects rather than broad implementation, comprehensive approaches will require greater preparation. As communities of practice like AI4LAM gather people for knowledge exchange and comparison of approaches at that project level, transitioning to broader adoption within an organization would benefit from a shared framework.

Even with community reporting, the parts of the work most often highlighted in these community presentations represent the outcomes of those projects and the methods employed. However the team and organizational dimensions are less frequently foregrounded, which leaves opaque essential methods for integrating subject matter expertise, staff competencies, and other critical considerations for the people involved in undertaking these projects.

This brief discussion of these challenges suggests that a shared framework may allow staff and leadership to take steps into practice, A shared framework might further present opportunities to experiment with intention, document data transformation and consequences, and suggest starting points to evaluate approaches for broader implementation.

# Lc Labs Exploring ML

Recent LC Labs initiatives have demonstrated the complexity inherent to benchmarking. Furthermore, it is imperative that the intersection of project and organizational objectives include opportunities to assess resources, collections, risk, and people encountered at each step. Committing to effectively centering people including users, staff and subjects of digitized items means that we must move with intention and integrate moments and mechanisms to ask critical questions of the approaches we are applying.

For the last several years, the LC Labs team has explored dimensions of machine learning through events, initiatives and experiments. We have hosted events, sponsored experiments and research, explored user needs from a range of angles, and frequently shared the outcomes of our work as part of our practice at LC Labs. We hosted a Machine Learning + Libraries Summit, alongside US- UK Digital Scholarship workshop in 2019 which also surfaced ML + crowdsourcing threads. From internal experimentation with Speech to Text Viewer to recommendations around socio-technical assessment and planning with the Intelligent Data Analytics report and a state of the field report on Machine Learning and Libraries; and from wildly successful and entertaining IIR experiments Newspaper Navigator and Citizen DJ, to the Collective Wisdom Project, Experimental Access initiative, and Humans in the Loop experiment, the LC Labs team and partners continue to investigate methods, models, and resources in context. Outcomes from this series of events and experiments have demonstrated that subject matter expertise is essential, that we must center approaches on humans and their real needs, and that we should experiment and iterate, while sharing outcomes [13].

Many of these endeavors were themselves informed by the work of the Digital Scholarship Working Group report [14]. Its foundational findings articulate essential needs for item-level metadata and rights assessment to enhance usability of digital collections - approaches that require human expertise and computational methods to address challenges of scale. Fundamentally, that work is iterative and woven together with many threads of collaboration and participation of colleagues.

These recurring recommendations have emerged from the ML-focused initiatives that LC Labs has sponsored:

* Cultivate responsible practices
* Develop appropriate solutions via iteration
* Make available training data for wider use
* Combine machine learning and crowdsourcing
* Sponsor interdisciplinary and interagency collaboration
* Support staff skills development
* Explore infrastructure, policy, and capacity

If and when implemented, these recommendations would benefit not only the Library of Congress but the wider library and archives field – so we continue to share them publicly via labs.loc.gov.

# Lc Labs Proposed Frameworks

The overarching themes from LC Labs experiments and reports on ML focus on developing a statement of values to guide decision-making around implementing AI in your organization, reinforcing that there is no one-size-fits-all solution when it comes to AI and LAMs and that the operationalizing of any kind of AI system will require more AI expertise across the organization. Building from these very important starting points, further frameworks are needed to help prioritize action and investment.

LC Labs AI and ML experiments have demonstrated and recommended several frameworks, including developing checklists, risk assessments and data archeologies, that encourage reflection and assessment of AI and ML goals against the capabilities and performance of existing models and data.

Practices are evolving across interdisciplinary sectors, accompanied by calls for implementation guidelines. Seeking useful examples that give structure to the community of practice and professional activity has surfaced tools and frameworks that offer practical and aspirational pathways to assess readiness, get started, think critically, and share practice, methods, tools, and insight. Examples include the NDSA Levels of Preservation, NIST AI Risk Framework, Collections as Data, Responsible Operations, and grant funded scholar-practitioner networks[15].

Additionally, methods of documenting datasets and models continue to be refined in interdisciplinary exchange [16] [17] [18]. AI model cards, for example, are lightweight documentation for AI models and are meant to support an informed decision about the use of a model by a non-expert, inspired by a nutrition fact label--you don't have to be a dietician to know the cautions around the food you’re eating. Model cards fit into a larger ecosystem of AI documentation. Documentation of the AI lifecycle helps support understanding, collaboration, sustainability, transparency, reusability. Components of an AI model card include [19]:

* Context: Express the intended user and use of the model, can also include what the model is not intended for
* Ethical Considerations - express the risks possible downstream considerations - environmental and for populations or groups, highlighted for non-technical stakeholders
* Data description: source, size of data and limitations of data (e.g. over 60% males represented)
* Quantitative analysis - overall quality of predictions in use cases

LC Labs developed an experimental framework to help rationalize what users and organizations have to benefit from specific ML or AI-enabled capabilities and to help gain insight into when and how to move toward implementation. The draft framework outlined below is for public comment, review and collaborative improvement.

## A: AI Capability Inventory and Assessment

In a spreadsheet, in column one, we are tracking types of ML or AI capabilities that have the potential to transform LAM digital services and categorizing them. The first broad category is divided between front of the house and back of the house services. Some capabilities are processes that are performed behind the curtain and then made public selectively, like an OCR process that helps to generate metadata to enhance search but is not displayed to users. Or, a process that creates one-second audio clips and sorts them by starting note so that they can be remixed and downloaded from an application. We categorized these capabilities as “enabling discovery at scale.” Additional capabilities in this category are generating granular metadata for items, pages, articles, and paragraphs to enhance search services, creating non-English language OCR, handwriting recognition, object classification, name entity identification and linking, and generating bibliographic data, among other tasks.

Another back of the house category is a group of tasks that are processes for local management and preservation of digital content and collections, we are calling these “enhanced collections processing and analysis.” Examples tasks here include using AI to assist in rights assessment of born-digital content, categorizing unstructured born-digital content like web archives and email, assisting in general document sorting for internal business processes, helping with inventory control systems, and creating data that feeds customizable presentations, exhibits and visualizations.

AI and ML can also be used by LAMs in the front of the house to further “augment and extend the user experience” by letting the public directly interface with AI-enabled services like recommending systems, text chat bots, voice recognition and answer systems, voice search services, or visual search tools.

LAM users are also employing AI tools themselves to analyze collections that are made available as data. This front of the house service we are calling “enabling research use” includes a wide range of processes a researcher would perform themselves, including corpus creation, technical methods research and network analysis, among others. The questions that have arisen with this area of capability, are around the surrounding reference services that would be required to support these uses in a responsible way.

In the rows of the spreadsheet we name the specific task or process being considered and in the further columns we capture aspects of the AI process that was examined. These are:

* user story,
* tools or methods tested,
* collection data utilized,
* benefits and risks for users, staff and the organization,
* evidence about the performance of the data or model,
* user or subject feedback and impact, and
* staff or training implications.

To try and summarize the assessment and to get at potential next steps in the exploration of a specific AI task or process, we developed a rough scoring system rating from one to five, one being the a process that could be closest to implementation.

1. Ready for large-scale implementation with guidelines.
2. Ready for small scale implementation with guidelines.
3. Build on current evidence and do more experimentation.
4. Design an initial experiment and engage stakeholders.
5. Identify and scope potential methods and services.

The evidence gathered through our experimentation to date points to the most potential for small-scale AI implementation in the ”enabling discovery at-scale” category, followed by the ‘enabling research use’ group of tasks. These are not surprising (or scientific) results because these are the categories we have done the most experimenting in. A broader set of use cases and feedback from other organizations testing this framework would be required to assess if this is a useful assessment model.

## B: A Data Processing Plan Template

One of the key lessons-learned from experiments involving machine learning or artificial intelligence is that characteristics of the data used to train models and how well it aligns to the target data (or data that will be processed with the model) directly indicates the quality of the model’s output. Most models are trained with contemporary born-digital data and don’t perform well when used to process historic or digitized content. The model and all data utilized in a processing task must be documented at each stage so the results can be analyzed--especially before implementing at scale. LC Labs developed a Data Processing Plan template as a starting point for a required set of documentation that technical staff, researchers, or vendors can compile before and after processing, transforming or generating any Library of Congress data. This documentation can help to ensure Library staff have more comprehensive information when deciding how to utilize data generated from experiments. The information will allow for responsible experimentation with Library of Congress data and the opportunity for Library staff to learn about how ML and AI can be effectively implemented.

The elements of the Data Processing Plan template are proposed below. The plan is a work in progress and is also shared with the goal of receiving feedback and community contribution. It is based on recommendations and existing data and algorithmic impact assessment guides. The goal for this plan would be to have staff, partners or vendors fill out an initial draft of the template for review and discussion. A final version of the template would then be compiled after the data had been processed. Each distinct data set that is used in an experiment would require a unique data processing plan.

Section A: General

Describing the goals of the experimental data processing or transformation, the scope of the intended workflow or pipeline, the data delivery format and specifications, and the description of the intended use of the generated data.

Section B: Data Documentation

Describing the data that will be processed, it’s title, technical composition, including file type, content type, number of items and relative size. The language of the dataset, the time period it covers, the genre and other description information about what intellectual content the dataset contains. Document any copyright, licensing, rights and/or privacy restrictions that could affect the Library’s (or the public’s) subsequent use of any data processed.

The relevant background context about the composition of the dataset. For example, a dataset may be organized as a single spreadsheet containing metadata about a collection or it may be a series of folders containing images derived from a particular source. The data’s provenance, or where it originated, how it was compiled, when, and by whom, and how the dataset is/was technically compiled, for example via an API query or bulk download.This section also covers the preprocessing steps. How has the dataset been classified, cleaned or otherwise prepared for the experiment? How was material selected for inclusion or exclusion in the dataset? Is the data organized according to a schema, content standard or other standards? If yes, which one?

Also document if there are any potential risks to people, communities and organizations if the dataset is used in the experiment and what are the strategies for risk mitigation. For example, searchable access to individual names and places could expose personal identifying information of private citizens. How will the experiment team mitigate these risks? For example, the team will select data that is over 125 years old to include in the experiment. How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?

Section C: When documenting a dataset for machine learning or artificial intelligence processes, describe the purpose of this dataset with relation to the ML/AL workflow. Explicitly address if it is being used as training, validation or test data. For training data, if the model is pre-trained, describe the data on which it was trained. If the model will be fine-tuned, outline the data involved in this process. If the model is being trained from scratch, outline the plan for creating training data. If *creating training data* using volunteers or paid participants (e.g. via crowdsourcing), please describe the workflow and incentive structure. If *validating training data* using volunteers or paid participants (e.g. via crowdsourcing), please describe the workflow and incentive structure. Document any known gaps in the dataset, such as missing instances or forms of representation. Address possible sources of bias in the dataset resulting from these discrepancies. Describe any steps taken to remediate or address gaps or bias in the dataset used in the ML/AI processing or the experiment overall.

Section D: Documentation of the model or models used, including the intended use of the model, the known limitations for the model and its copyright and licensing details. Before processing, document the predicted performance metrics of the model and after each stage of processing and fine tuning, document the actual performance metrics. Establish an audit schedule for how often and how many times the performance metrics will be checked and define a range of successful algorithmic performance. Draw a workflow or pipeline description and diagram, including plans for conducting annotation and validation process, including an overview of supervised or unsupervised machine learning passes.

# A Community Call to Action

With accessible and computable collections data, ML and AI methods can be used to enable discovery at scale, enhance collections processing and analysis, enable computational research and augment user experiences. This is the promise that has yet to be realized in LAMs. In sharing these frameworks, we want to continue a community discussion about developing structures that support informed decisions about emerging technologies in LAMs. Developing these initial assessments has been clarifying for prioritizing the next experiments in LC Labs and our hope is that a fuller set of use cases and input could make them useful for more organizations. We invite you to test it out and experiment with different use cases and designs and figure out what works and what does not work in your context. In the coming months, we will aim to come together again to continue iterating on these frameworks together. We are continually inspired by the work of our peers and colleagues and eager for feedback, particularly from the recently formed groups who will evaluate AI and ML practice in LAMs with a specific focus on equity and inclusive justice. The NDSA Levels of Digital Preservation are an excellent model for very actionable and digestible documentation. Extending this concept to AI and ML could help to ensure the informed and responsible adoption of these technologies across the LAM sector.

#### ACKNOWLEDGMENT

The authors wish to thank the researchers, collaborators, and peer community for participation in LC Labs research, experimentation, events, and outcomes. Furthermore, they wish to express gratitude to Library of Congress colleagues who have enabled exploration of uses and methods through their work to acquire, describe, support, manage, provide access to, and preserve collections.

# REFERENCES

1. M. Phillips, J. Bailey, A. Goethals, T. Owens. “The NDSA Levels of Digital Preservation : An Explanation and Uses.” Library of Congress. 2013. <https://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf>
2. The Data Nutrition Project. <https://datanutrition.org/>
3. T. Padilla. “*Responsible Operations: Data Science, Machine Learning, and AI in Libraries*.” Dublin, OH: OCLC Research. 2019. <https://doi.org/10.25333/xk7z-9g97>.
4. ExLibris. “Artificial Intelligence in the Library: Advantages, Challenges and Tradition.” 2018. <https://cdn2.hubspot.net/hubfs/2909474/Ex%20Libris%20Artificial%20Intelligence%20White%20Paper.pdf>
5. R. Cordell. “[Machine Learning + Libraries: A Report on the State of the Field](https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf).” Library of Congress. 2020. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf>
6. E. Lorang, L. Soh, Y. Liu, and C. Pack. Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project. Library of Congress. 2020. <https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf>
7. E.Jakeway, L. Algee, L. Allen, M.Ferriter, J. Mears, A. Potter, K. Zwaard. Machine Learning & Libraries Summit Event Summary. Library of Congress. 2019. [ML-Event-Summary-Final-2020-02-13.pdf (loc.gov)](https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf)
8. W. A. Ingram, S. Johnson. Ensuring SCholarly Access to Government Archives and Records. Virginia Tech University Libraries in partnership with Virginia Tech Center for Humanities and

the U.S. National Archives and Records Administration. 2021. <https://vtechworks.lib.vt.edu/handle/10919/108067>

1. AEOLIAN Network. <https://www.aeolian-network.net/>
2. The Collective Wisdom Project. <https://collectivewisdomproject.org.uk/>
3. Library of Congress Digital Strategy. 2018. <https://loc.gov/digital-strategy>
4. M. Vajcner. “The Importance of Context for Digitized Archival Collections.” The Importance of Context for Digitized Archival Collections, vol. 11, no.1, April 2008.
5. S. Averkamp, K. Willette, A. Rudersdorf, M. Ferriter. Humans-in-the-Loop Recommendations Report. Library of Congress. 2021. <https://labs.loc.gov/static/labs/work/reports/LC-Labs-Humans-in-the-Loop-Recommendations-Report-final.pdf>
6. A. Potter, G. Harris, K. Zwaard, D. Brunton, S. Garfinkel, J. Hessler, C. Maher, J, Mears, N. Saylor, S. Stillo, C.Townsend. Digital Scholarship at the Library of Congress: User demand, current practices, and options for expanded services. Library of Congress. 2020.
7. O. Murphy, E. Villaespesa. The Museums + AI Network. AI: A Museum Planning Toolkit. Goldsmiths, University of London. 2020.<https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf>
8. T. Gebru, J.Morgenstern, B. Vecchione, J. W. Vaughn, Hanna Wallach, H. Daume III, K. Crawford. “Datasheets for Datasets.” v8. 2021 <https://arxiv.org/abs/1803.09010>
9. M. Mitchell et al., “Model Cards for Model Reporting.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–29. 2019. <https://doi.org/10.1145/3287560.3287596>
10. E. M. Bender & B. Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics* 6: 587–604. 2018. <https://doi.org/10.1162/tacl_a_00041>.
11. B. Kopp. 2022. Framework presented as part of the General Services Administration Responsible AI Community of Practice presentation. Draft documents are only available to community members at the time of submission/publication.