Attachment J2 - Data Processing Plan Template

This template is provided to help partners and vendors understand the documentation and planning requirements for processing Library of Congress data in the context of experimentation and research. After an experiment is awarded and before any data processing tasks are performed, vendors and/or partners shall submit an initial draft of this template to the Library for review and discussion. A final version of the template shall be delivered after the data has been processed with all of the relevant information completed. Each distinct data set that is used in an experiment will require a unique data processing plan.

Section A: General (required)

A1: Goals of experiment. (consult Library/task order)
Fill in based on the Library of Congress Statement of Work or Task Order. The goals of the experiment are to help the Library answer the following research questions: How can the Library advance the outputs of the Exploring Computational Description task order (TO1 from the Digital Innovation IDIQ) to: A. refine quality standards and assessment methods for applying ML methods to generating specific MARC catalog fields, and B. use this information to develop workflows that combine several ML models or methods and human review by Library of Congress catalogers and digital collections staff? And, in particular, to identify: 1. Where are the most effective combinations of automation and human intervention in generating high-quality catalog records that will be usable at the Library of Congress? 2. What are the benefits, risks, and requirements for building a pilot application for ML-assisted cataloging workflows? The goal is, for this model, is to: ● measure the quality of the outputs (using standard metrics) ● gather any other additional data that can assist in the overall assessment of the benefits, risks, and costs to the Library as part of the reporting phase of the project ● evaluate the use of this model (or models) in a workflow that integrates human review by Library of Congress catalogers and digital collections staff The primary inputs to the experiment are in the form: ● of electronic publications (ebooks) as PDF and ePub, with accompanying ● MARC records (from MARCXML) and the primary expected outputs are:

- Generated catalog records for the *test* subset of the ebooks
- A record of any training parameters or other settings used in training the model
- Metrics which compare the generated catalog records to the actual catalog records for the same ebook (see attachment below)
- Exports of the data models generated (where possible or practical)
- Documentation for any code produced as part of the experiment (with the caveat that the code is intended primarily to test approaches to generating catalog metadata, rather than intended for production use)
- Testable user-facing prototypes that integrate the outputs of the model with a workflow suitable for use in a human-in-the-loop workflow by catalogers and other library staff

With this information to form part of the final report, synthesized with other information from desk research.

A2: Describe the scope of the intended workflow or pipeline. (consult Library/task order)

Fill in based on the Library of Congress Statement of Work or Task Order.

The intended workflow is to generate bibliographic metadata from ebooks in epub and PDF formats.

The goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible. The minimum fields to be generated are:

- titles,
- author names,
- unique identifiers,
- date of issuance,
- date of creation,
- genre/form, and
- subject terms.

In the case of this particular model, the expected scope is that the model will generate all core bibliographic metadata fields *other than LCSH subjects* and provide this data in a form suitable for further transformation in Marc records and for integration into a low-fidelity human-in-the-loop (HITL) prototype.

This model is unlikely to be well suited to LCSH subject classification given the very long tail of subject classifications in use, however, preliminary subjects may be generated for further downstream use in subsequent prototypes.

The broad aim for this experiment is to assess approaches to text and token classification on LoC ebooks which use open source Large Language Models or "generative AI" *without extensive fine-tuning or training of model weights*.

The second data processing plan (Model 2: LLM Fine Tuning) covers the fine-tuning of model weights using supervised learning.

We can think of this model as a broad general evaluation of the current state of the art open source LLMs and identification of the best approaches for utilizing these LLMs to produce catalog records by:

- basic prompting
- few-shot training (or in-context learning)
- constraining or structuring outputs

We expect to evaluate multiple open source Large Language Models, rather than a single Language Model

This particular workflow will focus primarily on the generation of core bibliographic metadata.

A3: Data delivery format and specifications for data generated in the experiment. (consult Library/task order)

Fill in based on the Library of Congress Statement of Work, Task Order or directive.

The primary output for this experiment will be:

- Bibliographic metadata fields for each ebook, with their accompanying labels.
- Generated genre terms for each ebook, with their accompanying identifiers in LCGFT.
- A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file.
- Exports of the data models generated (where possible).
- Documentation for any code produced as part of the experiment (with the caveat that the code is intended primarily to test approaches to generating catalog metadata, rather than intended for production use)
- Metrics which compare the generated catalog records to the actual catalog records for the same ebook (see attachment below). These will be provided as JSON files, and as CSV/XLSX files.
- A lightweight low-fidelity prototype which integrates the data in a user testable UI for further review for suitability in a human-in-the-loop cataloging workflow

A4: Description of intended use

Please describe how the data will be used in the experiment.

The experiment will direct one or more Large Language Models to use generative approaches to create catalog records. These directions may incorporate a number of existing catalog records as exemplars or for use in iteratively improving the quality of outputs via few-short or in-context learning and other methods.

Section B: Data Documentation (required)

Please fill out a complete chart *for each existing dataset under consideration for use in the experiment.* All experiments must have Sections A and B filled out. If the experiment involves machine learning or other artificial intelligence, Section B3 and Section C must also be filled out.

B1: Description of Dataset	
a) Title of dataset	*LCP Ebook dataset*
b) Composition 1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size. 2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains. 3. Please also include 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 dataset consists of ebooks and MarcXML files with catalog records for those ebooks. 1. Technical composition: a. Total number of items: 123778 i. 1777 duplicates ii. 119,823 unique ebooks b. File type: PDF and ePub. Approx. ⅓ of the files are PDFs and the remaining ⅔ are ePubs. c. Content type: ebooks d. Relative size: ~1TB 2. Full data audit to follow. a. Languages (35 languages): i. English ~120,000 records ii. Spanish ~1000 records iii. German ~700 records iv. Other: ~700 records b. Genre: Approx 6% of the records have a listed genre. For details see full data audit. c. Summary: Approx 57,000 records have publisher or other summaries d. Period: 21st century ebooks. For details see full data audit. 3. The dataset comprises four discrete sub-collections: a. CIP (1113,390 items) b. Open access (5835 items) c. E Deposit ebooks (403 items) d. Legal reports (3750 items) Each collection is organized as a folder of ebooks in PDF or ePub format. Accompanying each folder is a single MARCXML file containing the catalog records for each of the ebooks within that sub-collection.

c) Provenance 1. Where did the information in this dataset originate? Please include relevant links where possible. 2. Include any version information if available.	The information in this dataset originated from four collections of LoC ebooks: 1. Ebooks provided as part of CIP prepublication cataloging 2. Ebooks provided as part of E-Deposit registration 3. Ebooks provided as part of the Open Access ebooks program 4. Legal reports An additional CIP dataset was generated in December of 2023 for inclusion in this experiment. Further details to be provided by LoC.
d) Compilation methods 1. How is/was this dataset compiled, when, and by whom? 2. Please include technical details of how the dataset is/was compiled, e.g. loc.gov API query, bulk download.	1. The dataset was compiled by Library of Congress staff, including Lauren Seroka, on behalf of Caroline Saccucci and Abigail Potter (Lc Labs). 2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment and made available for further data for this experiment. 3. Additional data such as authority records, and additional MARCXML records were downloaded via the public downloads available via the MDSConnect program. Further details to be provided by LoC.
e) Preprocessing steps 1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment? 2. How was material selected for inclusion or exclusion in the dataset? 3. Is the data organized according to a schema, content standard, or other standard? If yes, which one?	1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were: a. Conversion of PDF and epub to plaintext using a mixture of tools, including Unstructured. b. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any)

	c. Normalize whitespace N.B. No other preparation is done before the experiment runs, as other "cleaning" steps such as stopword removal and lemmatization are specific to particular steps. 2. This question should be answered by LoC staff. In terms of the experiment, all ebooks will be used as part of the training, validation and test splits as long as the files are compatible. Exclusion will be for technical reasons only (invalid PDF or ePub file). 3. The metadata is organized as MARCXML files following usual LoC cataloging practice.
f) Potential risks to people, communities and organizations & strategies for risk mitigation: 1. What potential risks or harms could result to people, communities and organizations from processing this dataset in the experiment? (For example: searchable access to individual names and places could expose personal identifying information of private citizens.) a. 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.)	
The experiment will not expose any of the data to the wider public, communities or organizations. The primary outputs will be metrics, and Marc records, which will be used for internal evaluation and assessment only. To the extent that there is a risk, the risk is primarily that material cataloged by automated processes may use potentially pejorative or dispreferred cataloging terms, for example, for subjects. However, this is only a risk to the extent that such terms both exist in the MarcXML provided and are part of the LCSH subject vocabulary.	
g) How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?	
As per f), the primary outputs of the project are metrics and sample catalog records. The materials are all modern ebooks with recent catalog records. The records will not, as part of this experiment, be made public. Catalog records will be made available to Library staff via the low-fidelity prototype UI, but no additional steps will be taken to address outdated or offensive items.	
h) Copyright, licensing, rights, and/or privacy restrictions	The material comprises a mixture of open access and copyrighted ebooks.

1. Describe in sufficient detail limitations on any intellectual property or privacy or other restrictions that will affect the Library's (or the public's) subsequent use of any data processed.	The project will not make public any of the ebooks or the metadata generated from these ebooks except with the prior consent and explicit authorisation of the Library.

Will the dataset be used in conjunction with machine learning or artificial intelligence processes? If yes, please fill out all of section C and section D.

Section C: Documentation of a dataset for machine learning or artificial intelligence processes

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.
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 randomly assigned examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset. We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. However, the overall volume of non-English material is low, so this may not be required. A small subset of the training data split will be used for few-shot learning, or prompt tuning.
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.
Each of the large language models that might be evaluated has been trained on its own dataset, and in some cases, the precise details of the training dataset is left unclear or deliberately held back for competitive advantage. In some cases, models may potentially be trained on copyright or non-public-domain information. However, the broad datasets tend to be the same for most models. For example, LLama-2 is trained on (information from wikidata): ● Webpages scraped by CommonCrawl ● Open source repositories of source code from GitHub ● Wikipedia in 20 different languages ● Public domain books from Project Gutenberg ● The LaTeX source code for scientific papers uploaded to ArXiv ● Questions and answers from Stack Exchange websites

and additionally fine-tuned using 27,540 prompt-response pairs created for Llama-2 and reinforcement learning with human feedback (RLHF) was used with a combination of 1,418,091 Meta examples and seven smaller datasets.

Similarly, Google say, for Gemma, that:

> These models were trained on a dataset of text data that includes a wide variety of sources, totaling 6 trillion tokens. Here are the key components:
>
> Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content.
> Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions.
> Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.

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

N/A

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

N/A

e) 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.

1. Describe any steps taken to remediate or address gaps or bias in a dataset used in the ML/AI processing or the experiment overall.

While the underlying datasets used in the training of foundation model LLMs (large language models) are a known source of bias, 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, rather than generating new text, 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).

Section D: Documentation of ML model (required for experiments involving machine learning or artificial intelligence)

All experiments involving machine learning or artificial intelligence must complete the chart below for *any* models under consideration for use in the experiment.

C1: Machine Learning or Artificial Intelligence Model	
a) Model Details	Model 1: LLM Prompting The primary purpose of this model is to test non-commercial or free-to-use Large Language Models (LLMs). We expect to potentially test multiple models such as: ● Llama2 ● Mistral / Mixtral ● Gemma ● Yi ● Qwen
b) Intended use	Automated extraction of bibliographic metadata from ebooks.
c) Limitations	-
d) Copyright and licensing details for the model	Varies depending on model. Yi, Qwen and Mistral use the Apache 2.0 license Meta Llama: https://llama.meta.com/faq#legal Gemma: https://ai.google.dev/gemma/terms
e) Link to documentation	Yi: https://arxiv.org/abs/2403.04652 Qwen: https://arxiv.org/abs/2309.16609 Llama2: https://llama.meta.com/llama2

	Google Gemma: https://blog.google/technology/developers/gemma-open-models/ Mistral: https://docs.mistral.ai/
f) Predicted performance metrics (range)	This experiment is partly designed to discover what the range of performance metrics might be, so this is currently unknown. From the data we will be gathering standard metrics as part of the process, including: ● Precision ● Recall ● F1-Score Other metrics can be derived from the confusion matrix as required (Kappa Score, Matthew's correlation class coefficient, etc) and we would expect to generate some of these at the end of the experiment(s) as we begin our final writeup.
g) Actual performance metrics	N/A - these will be gathered as part of the experiment
h) Audit schedule (how often and how many times will performance metrics be checked?)	We would expect to gather metrics once at the end of the training and evaluation cycle.
i) Definitions of successful algorithmic performance. Specifically, performance evaluation factors and accuracy and performance results at each stage of the workflow and for each overall pass through the pipeline.	
Part of this experiment is to assist the LoC to determine what counts as successful algorithmic performance. For example, general figures such as 70% accuracy are sometimes used to define "ideal" or "good" model performance. For some, but not all fields this is probably an unreasonably high threshold. Previous experiments (Task Order One) approached scores of between 80 and 90% for some fields in real-world scenarios. We would expect similar or better figures for some, but not all fields. Subject or genre fields may have much lower levels of performance. Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report.	
i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning.	

We will, for each LLM we test, we will:

1. Take plaintext extracted from PDFs and ePubs stored in an Amazon S3 bucket
2. Create training files (exact formats and specifications vary between models and approaches).
    a. Generally, we will need to extract sequences from the full text that match the relevant sections of the document
    b. Tag these sequences with their label (Title, ISBN, LCCN, etc)
    c. Format these into the right corpus/training format
3. Create prompts for information extraction
4. Evaluate the prompts against the dataset
5. Refine and improve the prompts to improve performance on the standard metrics

Testing and evaluating prompts will be integrated into an MLFlow based pipeline. However, we will use additional tools, such as LangChain, Stanford DSpy, Guidance and others to test and evaluate the workflows and improve the outputs.

Note, that in this initial high level evaluation,we will be using:

- prompting
- in context learning
- few-shot learning

along with tools that can iteratively improve the outputs from workflows of this type.

We would *not* expect to be fine-tuning the models at this stage. A prompt-first approach is computationally efficient, and can be run with relatively small numbers of documents to generate optimal or near-optimal prompts for extracting bibliographic metadata from documents.

We can use this approach to get a good baseline for model performance and to identify the best candidate models for fine-tuning in Model 2: LLM Fine-tuning.

After each of the models is complete and tested we will select the best performing model or models.

For the best performing model we will:

6. Create an “automatic cataloging” workflow pipeline for all of the core Marc fields in scope for the Task Order. Taking, for each field, the highest scoring model if more than one model can produce the same data.
7. Run the workflow across the *Test* set of ebooks (not used in training and refinement) to generate catalog data for each ebook: we would expect, at this stage, that these would take the form of lightweight JSON files that we can serialize to Marc or to BibFrame later, as required.
8. Generate metrics (F-Score, Precision, etc) for each of the primary field types in the records.
9. Store the generated catalog data and metrics for use in the final report

The infrastructure will comprise:

- One or more AWS EC2 instances with GPU processors and fast storage for model training and testing
- Amazon AWS S3 buckets for:
    - PDFs, ePubs and MarcXML files (as provided by LoC)
    - Project configuration, plaintext files, etc.
- MLFLow and Ray for infrastructure and experiment management

Attachment J2 - Data Processing Plan Template

This template is provided to help partners and vendors understand the documentation and planning requirements for processing Library of Congress data in the context of experimentation and research. After an experiment is awarded and before any data processing tasks are performed, vendors and/or partners shall submit an initial draft of this template to the Library for review and discussion. A final version of the template shall be delivered after the data has been processed with all of the relevant information completed. Each distinct data set that is used in an experiment will require a unique data processing plan.

Section A: General (required)

A1: Goals of experiment. (consult Library/task order)
Fill in based on the Library of Congress Statement of Work or Task Order. The goals of the experiment are to help the Library answer the following research questions: How can the Library advance the outputs of the Exploring Computational Description task order (TO1 from the Digital Innovation IDIQ) to: A. refine quality standards and assessment methods for applying ML methods to generating specific MARC catalog fields, and B. use this information to develop workflows that combine several ML models or methods and human review by Library of Congress catalogers and digital collections staff? And, in particular, to identify: 1. Where are the most effective combinations of automation and human intervention in generating high-quality catalog records that will be usable at the Library of Congress? 2. What are the benefits, risks, and requirements for building a pilot application for ML-assisted cataloging workflows? The goal is, for this model, is to: ● measure the quality of the outputs (using standard metrics) ● gather any other additional data that can assist in the overall assessment of the benefits, risks, and costs to the Library as part of the reporting phase of the project ● evaluate the use of this model (or models) in a workflow that integrates human review by Library of Congress catalogers and digital collections staff The primary inputs to the experiment are in the form: ● of electronic publications (ebooks) as PDF and ePub, with accompanying ● MARC records (from MARCXML) and the primary expected outputs are:

- Generated catalog records for the *test* subset of the ebooks
- A record of any training parameters or other settings used in training the model
- Metrics which compare the generated catalog records to the actual catalog records for the same ebook (see attachment below)
- Exports of the data models generated (where possible or practical)
- Documentation for any code produced as part of the experiment (with the caveat that the code is intended primarily to test approaches to generating catalog metadata, rather than intended for production use)
- Testable user-facing prototypes that integrate the outputs of the model with a workflow suitable for use in a human-in-the-loop workflow by catalogers and other library staff

With this information to form part of the final report, synthesized with other information from desk research.

A2: Describe the scope of the intended workflow or pipeline. (consult Library/task order)

Fill in based on the Library of Congress Statement of Work or Task Order.

The intended workflow is to generate bibliographic metadata from ebooks in epub and PDF formats.

The goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible. The minimum fields to be generated are:

- titles,
- author names,
- unique identifiers,
- date of issuance,
- date of creation,
- genre/form, and
- subject terms.

In the case of this particular model, the expected scope is that the model will generate all core bibliographic metadata fields *other than LCSH subjects* and provide this data in a form suitable for further transformation in Marc records and for integration into a low-fidelity human-in-the-loop (HITL) prototype.

This model is unlikely to be well suited to LCSH subject classification given the very long tail of subject classifications in use, however, preliminary subjects may be generated for further downstream use in subsequent prototypes.

The broad aim for this experiment is to fine-tuning model weights using supervised learning on the highest performing model or models identified through the first experiment: **Model 1: LLM Prompting / few-shot.**

This particular workflow will focus primarily on the generation of core bibliographic metadata whereas other workflows and the use of supervised learning to fine-tune an existing foundational LLM using existing catalog records to improve performance beyond prompt-engineering and in-context learning.

A3: Data delivery format and specifications for data generated in the experiment. (consult Library/task order)
Fill in based on the Library of Congress Statement of Work, Task Order or directive. The primary output for this experiment will be: ● Bibliographic metadata fields for each ebook, with their accompanying labels. ● Generated genre terms for each ebook, with their accompanying identifiers in LCGFT. ● A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file. ● Exports of the data models generated (where possible). ● Documentation for any code produced as part of the experiment (with the caveat that the code is intended primarily to test approaches to generating catalog metadata, rather than intended for production use) ● Metrics which compare the generated catalog records to the actual catalog records for the same ebook (see attachment below). These will be provided as JSON files, and as CSV/XLSX files. ● A lightweight low-fidelity prototype which integrates the data in a user testable UI for further review for suitability in a human-in-the-loop cataloging workflow
A4: Description of intended use
Please describe how the data will be used in the experiment. The experiment will direct one or more Large Language Models to use generative approaches to create catalog records. These directions will incorporate a number of existing catalog records for use in fine-tuning the model through parameter efficient approaches to fine-tuning models.

Section B: Data Documentation (required)

Please fill out a complete chart *for each existing dataset under consideration for use in the experiment.* All experiments must have Sections A and B filled out. If the experiment involves machine learning or other artificial intelligence, Section B3 and Section C must also be filled out.

B1: Description of Dataset	
a) Title of dataset	*LCP Ebook dataset*
b) Composition	The dataset consists of ebooks and MarcXML files with catalog records for those ebooks.

1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size. 2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains. 3. Please also include 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.	1. Technical composition: a. Total number of items: 123778 i. 1777 duplicates ii. 119,823 unique ebooks a. File type: PDF and ePub. Approx. ⅓ of the files are PDFs and the remaining ⅔ are ePubs. b. Content type: ebooks c. Relative size: ~1TB 2. Full data audit to follow. a. Languages (35 languages): i. English ~120,000 records ii. Spanish ~1000 records iii. German ~700 records iv. Other: ~700 records b. Genre: Approx 6% of the records have a listed genre. For details see full data audit. c. Summary: Approx 57,000 records have publisher or other summaries d. Period: 21st century ebooks. For details see full data audit. 3. The dataset comprises four discrete sub-collections: a. CIP (1113,390 items) b. Open access (5835 items) c. E Deposit ebooks (403 items) d. Legal reports (3750 items) Each collection is organized as a folder of ebooks in PDF or ePub format. Accompanying each folder is a single MARCXML file containing the catalog records for each of the ebooks within that sub-collection.
c) Provenance 1. Where did the information in this dataset originate? Please include relevant links where possible.	The information in this dataset originated from four collections of LoC ebooks: 1. Ebooks provided as part of CIP prepublication cataloging

2. Include any version information if available.	2. Ebooks provided as part of E-Deposit registration 3. Ebooks provided as part of the Open Access ebooks program 4. Legal reports An additional CIP dataset was generated in December of 2023 for inclusion in this experiment. Further details to be provided by LoC.
d) Compilation methods 1. How is/was this dataset compiled, when, and by whom? 2. Please include technical details of how the dataset is/was compiled, e.g. loc.gov API query, bulk download.	1. The dataset was compiled by Library of Congress staff, including Lauren Seroka, on behalf of Caroline Saccucci and Abigail Potter (Lc Labs). 2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment and made available for further data for this experiment. 3. Additional data such as authority records, and additional MARCXML records were downloaded via the public downloads available via the MDSConnect program. Further details to be provided by LoC.
e) Preprocessing steps 1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment? 2. How was material selected for inclusion or exclusion in the dataset? 3. Is the data organized according to a schema, content standard, or other standard? If yes, which one?	1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were: a. Conversion of PDF and epub to plaintext using a mixture of tools, including Unstructured. b. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any) c. Normalize whitespace N.B. No other preparation is done before the experiment runs, as other "cleaning" steps such as stopword removal and lemmatization are specific to particular steps. 2. This question should be answered by LoC staff. In terms of the experiment, all ebooks will be

	used as part of the training, validation and test splits as long as the files are compatible. Exclusion will be for technical reasons only (invalid PDF or ePub file). 3. The metadata is organized as MARCXML files following usual LoC cataloging practice.
f) Potential risks to people, communities and organizations & strategies for risk mitigation: 1. What potential risks or harms could result to people, communities and organizations from processing this dataset in the experiment? (For example: searchable access to individual names and places could expose personal identifying information of private citizens.) a. 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.)	
The experiment will not expose any of the data to the wider public, communities or organizations. The primary outputs will be metrics, and Marc records, which will be used for internal evaluation and assessment only. To the extent that there is a risk, the risk is primarily that material cataloged by automated processes may use potentially pejorative or dispreferred cataloging terms, for example, for subjects. However, this is only a risk to the extent that such terms both exist in the MarcXML provided and are part of the LCSH subject vocabulary.	
g) How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?	
As per f), the primary outputs of the project are metrics and sample catalog records. The materials are all modern ebooks with recent catalog records. The records will not, as part of this experiment, be made public. Catalog records will be made available to Library staff via the low-fidelity prototype UI, but no additional steps will be taken to address outdated or offensive items.	
h) Copyright, licensing, rights, and/or privacy restrictions 1. Describe in sufficient detail limitations on any intellectual property or privacy or other restrictions that will affect the Library's (or the public's) subsequent use of any data processed.	The material comprises a mixture of open access and copyrighted ebooks. The project will not make public any of the ebooks or the metadata generated from these ebooks except with the prior consent and explicit authorisation of the Library.

Will the dataset be used in conjunction with machine learning or artificial intelligence processes? If yes, please fill out all of section C and section D.

Section C: Documentation of a dataset for machine learning or artificial intelligence processes

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.
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 randomly assigned examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset. We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. However, the overall volume of non-English material is low, so this may not be required. A small subset of the training data split will be used for few-shot learning, or prompt tuning.
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.
Each of the large language models that might be evaluated has been trained on its own dataset, and in some cases, the precise details of the training dataset is left unclear or deliberately held back for competitive advantage. In some cases, models may potentially be trained on copyright or non-public-domain information. However, the broad datasets tend to be the same for most models. For example, LLama-2 is trained on (information from wikidata): ● Webpages scraped by CommonCrawl ● Open source repositories of source code from GitHub ● Wikipedia in 20 different languages ● Public domain books from Project Gutenberg ● The LaTeX source code for scientific papers uploaded to ArXiv ● Questions and answers from Stack Exchange websites and additionally fine-tuned using 27,540 prompt-response pairs created for Llama-2 and reinforcement learning with human feedback (RLHF) was used with a combination of 1,418,091 Meta examples and seven smaller datasets. Similarly, Google say, for Gemma, that: > These models were trained on a dataset of text data that includes a wide variety of sources, totaling 6 trillion tokens. Here are the key components: > Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content.

Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions. Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.
c) If creating training data using **volunteers or paid participants (e.g. via crowdsourcing),** please describe the workflow and incentive structure.
N/A
d) If validating training data using **volunteers or paid participants (e.g. via crowdsourcing),** please describe the workflow and incentive structure.
N/A
e) 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. 1. Describe any steps taken to remediate or address gaps or bias in a dataset used in the ML/AI processing or the experiment overall.
While the underlying datasets used in the training of foundation model LLMs (large language models) are a known source of bias, 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, rather than generating new text, 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).

Section D: Documentation of ML model (required for experiments involving machine learning or artificial intelligence)

All experiments involving machine learning or artificial intelligence must complete the chart below for *any* models under consideration for use in the experiment.

C1: Machine Learning or Artificial Intelligence Model	
a) Model Details	Model 2: LLM Fine-tuning The primary purpose of this model is to test non-commercial or free-to-use Large Language Models (LLMs). We expect to potentially train/fine-tune a single LLM identified through Model 1: LLM Prompting which this model will be taken from ● Llama2 ● Mistral / Mixtral ● Gemma ● Yi ● Qwen
b) Intended use	Automated extraction of bibliographic metadata from ebooks.
c) Limitations	-
d) Copyright and licensing details for the model	Varies depending on model. Yi, Qwen and Mistral use the Apache 2.0 license Meta Llama: https://llama.meta.com/faq#legal Gemma: https://ai.google.dev/gemma/terms
e) Link to documentation	Yi: https://arxiv.org/abs/2403.04652 Qwen: https://arxiv.org/abs/2309.16609 Llama2: https://llama.meta.com/llama2 Google Gemma: https://blog.google/technology/developers/gemma-open-models/ Mistral: https://docs.mistral.ai/
f) Predicted performance metrics (range)	This experiment is partly designed to discover what the range of performance metrics might be, so this is currently unknown. From the data we will be gathering standard metrics as part of the process, including:

	• Precision • Recall • F1-Score Other metrics can be derived from the confusion matrix as required (Kappa Score, Matthew's correlation class coefficient, etc) and we would expect to generate some of these at the end of the experiment(s) as we begin our final writeup.
g) Actual performance metrics	N/A - these will be gathered as part of the experiment
h) Audit schedule (how often and how many times will performance metrics be checked?)	We would expect to gather metrics once at the end of the training and evaluation cycle.
i) Definitions of successful algorithmic performance. Specifically, performance evaluation factors and accuracy and performance results at each stage of the workflow and for each overall pass through the pipeline.	
Part of this experiment is to assist the LoC to determine what counts as successful algorithmic performance. For example, general figures such as 70% accuracy are sometimes used to define "ideal" or "good" model performance. For some, but not all fields this is probably an unreasonably high threshold. Previous experiments (Task Order One) approached scores of between 80 and 90% for some fields in real-world scenarios. We would expect similar or better figures for some, but not all fields. Subject or genre fields may have much lower levels of performance. Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report.	
i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning.	

We would expect Model 1: LLM Prompting to have identified:

1. A good candidate base model to fine-tune
2. Effective prompts which generate usable bibliographic metadata
3. Any methods for constraining the outputs of the LLM to a particular format or schema

For this workflow, we will test a number of parameter-efficient fine-tuning methods (PEFT) which will generate additional model weights by fine-tuning models using a dataset or datasets derived from the canonical catalog records from the MARCXML.

We will use one or more frameworks such as:

- Llama-Factory
- Hugging Face PEFT
- OpenAccess AI Collective axolotl

To generate new model weights using fine-tuning.

For the best performing model we will:

1. Merge the new model weights to create a custom fine-tuned language model for MARC cataloging
2. Create an “automatic cataloging” workflow pipeline for all of the core Marc fields in scope for the Task Order.
3. Run the workflow across the *Test* set of ebooks (not used in training and refinement) to generate catalog data for each ebook: we would expect, at this stage, that these would take the form of lightweight JSON files that we can serialize to Marc or to BibFrame later, as required.
4. Generate metrics (F-Score, Precision, etc) for each of the primary field types in the records.
5. Store the generated catalog data and metrics for use in the final report

The infrastructure will comprise:

- One or more AWS EC2 instances with GPU processors and fast storage for model training and testing
- Amazon AWS S3 buckets for:
    - PDFs, ePubs and MarcXML files (as provided by LoC)
    - Project configuration, plaintext files, etc.
- MLFLow and Ray for infrastructure and experiment management
- Additional libraries and tooling to facilitate fine-tuning

Attachment J2 - Data Processing Plan Template

This template is provided to help partners and vendors understand the documentation and planning requirements for processing Library of Congress data in the context of experimentation and research. After an experiment is awarded and before any data processing tasks are performed, vendors and/or partners shall submit an initial draft of this template to the Library for review and discussion. A final version of the template shall be delivered after the data has been processed with all of the relevant information completed. Each distinct data set that is used in an experiment will require a unique data processing plan.

Section A: General (required)

A1: Goals of experiment. (consult Library/task order)
Fill in based on the Library of Congress Statement of Work or Task Order. The goals of the experiment are to help the Library answer the following research questions: How can the Library advance the outputs of the Exploring Computational Description task order (TO1 from the Digital Innovation IDIQ) to: A. refine quality standards and assessment methods for applying ML methods to generating specific MARC catalog fields, and B. use this information to develop workflows that combine several ML models or methods and human review by Library of Congress catalogers and digital collections staff? And, in particular, to identify: 1. Where are the most effective combinations of automation and human intervention in generating high-quality catalog records that will be usable at the Library of Congress? 2. What are the benefits, risks, and requirements for building a pilot application for ML-assisted cataloging workflows? The goal is, for this model, is to: ● measure the quality of the outputs (using standard metrics) ● gather any other additional data that can assist in the overall assessment of the benefits, risks, and costs to the Library as part of the reporting phase of the project ● evaluate the use of this model (or models) in a workflow that integrates human review by Library of Congress catalogers and digital collections staff The primary inputs to the experiment are in the form: ● of electronic publications (ebooks) as PDF and ePub, with accompanying ● MARC records (from MARCXML) and the primary expected outputs are:

- Generated catalog records for the *test* subset of the ebooks
- A record of any training parameters or other settings used in training the model
- Metrics which compare the generated catalog records to the actual catalog records for the same ebook (see attachment below)
- Exports of the data models generated (where possible or practical)
- Documentation for any code produced as part of the experiment (with the caveat that the code is intended primarily to test approaches to generating catalog metadata, rather than intended for production use)
- Testable user-facing prototypes that integrate the outputs of the model with a workflow suitable for use in a human-in-the-loop workflow by catalogers and other library staff

With this information to form part of the final report, synthesized with other information from desk research.

A2: Describe the scope of the intended workflow or pipeline. (consult Library/task order)

Fill in based on the Library of Congress Statement of Work or Task Order.

The intended workflow is to generate bibliographic metadata from ebooks in epub and PDF formats.

The goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible.

In the case of this particular model, the expected scope is *not* that the model will generate all core bibliographic metadata fields directly.

Rather, this model will focus on:

- matching machine generated data with authority records for:
  - creators
  - subjects
- matching machine generated data with other catalog records to identify *related records* to aid in:
  - subject classification
  - authority identification
- generating additional data such as summaries for human review

The broad aim for this experiment is take ML generated catalog metadata from Models 1 and 2 and use this metadata to retrieve related data:

- other catalog records for related books
- relevant authority records for personal names or subject

and present this data to the cataloger along-side the machine generated data to assist them in their cataloging workflow.

A3: Data delivery format and specifications for data generated in the experiment. (consult Library/task order)

Fill in based on the Library of Congress Statement of Work, Task Order or directive. The primary output for this experiment will be: ● Workflows for retrieving candidate subjects and personal names from LCSH and LCNAF ● Workflows for retrieving catalog records for related works ● Static samples of data for a relevant test set or evaluation set of records ● A lightweight low-fidelity prototype which integrates the data in a user testable UI for further review for suitability in a human-in-the-loop cataloging workflow
A4: Description of intended use
Please describe how the data will be used in the experiment. -

Section B: Data Documentation (required)

Please fill out a complete chart *for each existing dataset under consideration for use in the experiment.* All experiments must have Sections A and B filled out. If the experiment involves machine learning or other artificial intelligence, Section B3 and Section C must also be filled out.

B1: Description of Dataset	
a) Title of dataset	*LCP Ebook dataset*
b) Composition 1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size. 2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains. 3. Please also include 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 dataset consists of ebooks and MarcXML files with catalog records for those ebooks. 1. Technical composition: a. Total number of items: 123778 i. 1777 duplicates ii. 119,823 unique ebooks a. File type: PDF and ePub. Approx. ⅓ of the files are PDFs and the remaining ⅔ are ePubs. b. Content type: ebooks c. Relative size: ~1TB 2. Full data audit to follow. a. Languages (35 languages): i. English ~120,000 records ii. Spanish ~1000 records iii. German ~700 records

	iv. Other: ~700 records b. Genre: Approx 6% of the records have a listed genre. For details see full data audit. c. Summary: Approx 57,000 records have publisher or other summaries d. Period: 21st century ebooks. For details see full data audit. 3. The dataset comprises four discrete sub-collections: a. CIP (1113,390 items) b. Open access (5835 items) c. E Deposit ebooks (403 items) d. Legal reports (3750 items) Each collection is organized as a folder of ebooks in PDF or ePub format. Accompanying each folder is a single MARCXML file containing the catalog records for each of the ebooks within that sub-collection.
c) Provenance 1. Where did the information in this dataset originate? Please include relevant links where possible. 2. Include any version information if available.	The information in this dataset originated from four collections of LoC ebooks: 1. Ebooks provided as part of CIP prepublication cataloging 2. Ebooks provided as part of E-Deposit registration 3. Ebooks provided as part of the Open Access ebooks program 4. Legal reports An additional CIP dataset was generated in December of 2023 for inclusion in this experiment. Further details to be provided by LoC.
d) Compilation methods	1. The dataset was compiled by Library of Congress staff, including Lauren Seroka,

1. How is/was this dataset compiled, when, and by whom? 2. Please include technical details of how the dataset is/was compiled, e.g. loc.gov API query, bulk download.	on behalf of Caroline Saccucci and Abigail Potter (Lc Labs). 2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment and made available for further data for this experiment. 3. Additional data such as authority records, and additional MARCXML records were downloaded via the public downloads available via the MDSConnect program. Further details to be provided by LoC.
e) Preprocessing steps 1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment? 2. How was material selected for inclusion or exclusion in the dataset? 3. Is the data organized according to a schema, content standard, or other standard? If yes, which one?	1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were: a. Conversion of PDF and epub to plaintext using a mixture of tools, including Unstructured. b. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any) c. Normalize whitespace N.B. No other preparation is done before the experiment runs, as other "cleaning" steps such as stopword removal and lemmatization are specific to particular steps. 2. This question should be answered by LoC staff. In terms of the experiment, all ebooks will be used as part of the training, validation and test splits as long as the files are compatible. Exclusion will be for technical reasons only (invalid PDF or ePub file). 3. The metadata is organized as MARCXML files following usual LoC cataloging practice.
f) Potential risks to people, communities and organizations & strategies for risk mitigation: 1. What potential risks or harms could result to people, communities and organizations from processing this dataset in the experiment? (For example: searchable access to individual names and places could expose personal identifying information of private citizens.)	

a. 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.)
The experiment will not expose any of the data to the wider public, communities or organizations. The primary outputs will be metrics, and Marc records, which will be used for internal evaluation and assessment only. To the extent that there is a risk, the risk is primarily that material cataloged by automated processes may use potentially pejorative or dispreferred cataloging terms, for example, for subjects. However, this is only a risk to the extent that such terms both exist in the MarcXML provided and are part of the LCSH subject vocabulary.
g) How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?
As per f), the primary outputs of the project are metrics and sample catalog records. The materials are all modern ebooks with recent catalog records. The records will not, as part of this experiment, be made public. Catalog records will be made available to Library staff via the low-fidelity prototype UI, but no additional steps will be taken to address outdated or offensive items.

h) Copyright, licensing, rights, and/or privacy restrictions 1. Describe in sufficient detail limitations on any intellectual property or privacy or other restrictions that will affect the Library's (or the public's) subsequent use of any data processed.	The material comprises a mixture of open access and copyrighted ebooks. The project will not make public any of the ebooks or the metadata generated from these ebooks except with the prior consent and explicit authorisation of the Library.

Will the dataset be used in conjunction with machine learning or artificial intelligence processes? If yes, please fill out all of section C and section D.

Section C: Documentation of a dataset for machine learning or artificial intelligence processes

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.

In this particular experiment, the dataset will be vectorized and indexed into a database for retrieval by LLMs or by other methods as part of the experiment.

However, the data will not be used in training, or validation as this particular workflow will not require any active training of new models.

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.

The most likely models that will be used as part of this experiment will be drawn from the Sentence Transformer family of models used to *embed* text as a series of high dimensional vectors.

A list of these models can be found:

https://www.sbert.net/docs/pretrained_models.html

Those models typically provide links to their training data on their model card, for example, MPNet at

https://huggingface.co/sentence-transformers/all-mpnet-base-v2#training-data

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

N/A

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

N/A

e) 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.

1. Describe any steps taken to remediate or address gaps or bias in a dataset used in the ML/AI processing or the experiment overall.

While the underlying datasets used in the training of foundation model LLMs (large language models) are a known source of bias, 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, rather than generating new text, 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).

Section D: Documentation of ML model (required for experiments involving machine learning or artificial intelligence)

All experiments involving machine learning or artificial intelligence must complete the chart below for *any* models under consideration for use in the experiment.

C1: Machine Learning or Artificial Intelligence Model	
a) Model Details	Model 3: Retrieval Augmented Generation and Vector DBs
b) Intended use	Automated extraction of bibliographic metadata from ebooks and related records from catalog corpora and authority records
c) Limitations	-
d) Copyright and licensing details for the model	Varies depending on model.
e) Link to documentation	https://www.sbert.net/index.html
f) Predicted performance metrics (range)	This experiment is partly designed to discover what the range of performance metrics might be, so this is currently unknown. From the data we will be gathering standard metrics as part of the process, including: ● Precision ● Recall ● F1-Score

	Other metrics can be derived from the confusion matrix as required (Kappa Score, Matthew's correlation class coefficient, etc) and we would expect to generate some of these at the end of the experiment(s) as we begin our final writeup.
g) Actual performance metrics	N/A - these will be gathered as part of the experiment
h) Audit schedule (how often and how many times will performance metrics be checked?)	We would expect to gather metrics once at the end of the training and evaluation cycle.

i) Definitions of successful algorithmic performance. Specifically, performance evaluation factors and accuracy and performance results at each stage of the workflow and for each overall pass through the pipeline.

Part of this experiment is to assist the LoC to determine what counts as successful algorithmic performance.

For example, general figures such as 70% accuracy are sometimes used to define "ideal" or "good" model performance. For some, but not all fields this is probably an unreasonably high threshold.

Previous experiments (Task Order One) approached scores of between 80 and 90% for some fields in real-world scenarios. We would expect similar or better figures for some, but not all fields. Subject or genre fields may have much lower levels of performance.

Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report.

i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning.

This experiment is designed to:

1. Take existing Library of Congress:
    a. authority data from LCNAF and LCSH
    b. catalog records from MDSConnect
2. Vectorize the data to create embeddings that can be stored in a vector database (such as QDrant or Weaviate) using either a Sentence Transformer model, or the same embedding model as the LLM we fine-tuned in Model 2.
3. Index the vectorized data

This data can be used in a number of ways:

4. To provide nearest neighbor matches for ML generated catalog records so that
    a. this model returns similar MARC records to the ML generated MARC record
    b. these related records can then be used by catalogers in a UI to select relevant subjects or authorities to include in an updated HITL improved version of the ML record

5. Provide nearest neighbor matches for LCSH subjects or LCNAF authorities so that candidate subjects, authors, editors, etc can be matched with relevant authority records without requiring manual search
6. Providing additional material to add to the processing context of the LLM trained in Models 1 and 2 (https://arxiv.org/pdf/2005.11401.pdf and https://research.ibm.com/blog/retrieval-augmented-generation-RAG) as part of a RAG workflow where the LLM will answer questions about the ebook being cataloged with the assistance of related record data and authority data retrieved from the database

For testing, we will:

1. Create an "automatic cataloging" workflow pipeline which takes as an input the LLM generated catalog metadata
2. Run the workflow across the *Test* set of ebooks (not used in training and refinement) to generate candidate related records and authorities to include in an
3. HITL cataloging UI
4. Evaluate the quality of this data via manual review during user testing of the final prototype

The infrastructure will comprise:

- One or more AWS EC2 instances with GPU processors and fast storage for model training and testing
- Amazon AWS S3 buckets for:
    - PDFs, ePubs and MarcXML files (as provided by LoC)
    - Project configuration, plaintext files, etc.
- MLFLow and Ray for infrastructure and experiment management