There are five models /plans:

1. GROBID (GeneRation Of BIbliographic Data) - a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications. The extraction includes bibliographical information (e.g. title, abstract, authors, affiliations, keywords) along with the text and document structure. We will use GROBID to provide some initial benchmarking with an off the shelf tool, and then train the model to be more tailored to LoC data. GROBID is also useful for generating XML files that we can use for text input for the following experiments.
2. Annif: automated subject indexing toolkit - a library from the National LIbrary of Finland which is designed for automated subject cataloging. Annif provides access to multiple different ML backends, so we can trial multiple different ML models and approaches (TF-IDF, MLLM, etc) and benchmark a wide range of approaches to subject and genre cataloging.
3. LoC Spacy: Spacy (spaCy · Industrial-strength Natural Language Processing in Python) with additional pipeline steps for LoC catalog metadata. Spacy is an industry standard NLP library, with extensive abilities to be trained and customized, and which we can use for the full range of metadata for the experiment, including subjects, genres and bibliographic metadata.
4. BERT: testing and training a wide range of BERT-derived large language models (BERT, RoBERTa, distilBERT, etc.) and transformer based approaches for token classification (titles, authors, dates, etc) and for text classification (subjects and genres).
5. NLP with Layout features: supplementing either 4 or 5 (depending on the outputs of the earlier experiments) with layout data such as page position, text size, text location, page number, recto/verso, etc in order to identify if visual information can add additional weighting/quality to the NLP models and to further refine data extraction for titles, authors, and other fields that have distinct positions, or formatting within the document.

# 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:

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

This particular data processing plan concerns the testing of machine learning models (or other computational approaches) for generating cataloging metadata from Library ebooks.

The goal is, for this model, to:

- measure the quality of the outputs (using some standard metrics)
- measure the cost (in terms of hours of person time, and in terms of compute costs)
- 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

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 hyperparameters 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)

- 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)

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.

While the goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible, with the minimum fields to be generated comprising:

- Titles
- Author names,
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form
- Subject terms.

It is not always the case, however, that each individual ML model is going to be suited to all of these.

Models that are suited to, for example, extreme multi-label text classification—such as generating subject or genre terms—are not necessarily the same models that are suited to named entity extraction or token classification, such as identifying an ISBN, copyright statement,  or Title for an ebook. While some models under test may attempt to generate all of the above types of catalog metadata, in some cases they will not.

In the case of the trained models shipped with GROBID, the expected scope is that the model will generate:

- Titles
- Author Names
- Unique identifiers (DOI, PII, ISSN, ISBN etc.)
- Keywords
- Copyright

Subjects and Genres will not be produced by this experiment.

The broad aim for this experiment is to assess GROBID and CRF (Conditional random field) based approaches to text and token classification on LoC ebooks. The existing trained models can provide a number of the minimum fields for a bibliographic record, but as well as evaluation with these, we will train and evaluate the header-extraction model(s) with data from the LoC ebook dataset. This will include the extraction of additional bibliographic metadata not targeted by the pre-trained models.

**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.
- N.B. The interim data format will not be Marc but TEI/XML that we can convert into Marc later. We can also extract the plaintext from the TEI/XML for use in successive experiments.
- 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). GROBID provides means of saving models to disk.
- 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.

**A4: Description of intended use**

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

The experiment will evaluate performance of GROBID on ebook pdfs, and train models using the ebook pdf and MARC data, and the resulting models will be used to generate bibliographic metadata.

The primary intended use for the data generated is as part of the final report, rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive | The dataset consists of ebooks and MarcXML files with catalog records for those ebooks.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs |

| | |
|---|---|
| information about what intellectual content the dataset contains.<br>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. | c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP  (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br><br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br><br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance<br>1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | The information in this dataset originated from four collections of LoC ebooks:<br><br>1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br><br>Further details to be provided by LoC. |
| d) Compilation methods | 1. The dataset was compiled by Library of Congress staff, including Lauren Seroka, |

<table>
<tr><td>1. How is/was this dataset compiled, when, and by whom?<br>2. Please include technical details of how the dataset is/was compiled, e.g. loc.gov API query, bulk download.</td><td>on behalf of Caroline Saccucci and Abigail Potter (Lc Labs).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br><br>Further details to be provided by LoC.</td></tr>
<tr><td>e) Preprocessing steps<br>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>3. Is the data organized according to a schema, content standard, or other standard? If yes, which one?</td><td>1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were:<br>a. Conversion of epub to PDF using e.g. Pandoc or Calibre.<br>b. Conversion of MARCXML records to a TEI format for training.<br><br>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)..<br><br>3. The metadata is organized as MarcXML files following usual LoC cataloging practice.</td></tr>
<tr><td colspan="2">f) Potential risks to people, communities and organizations &amp; strategies for risk mitigation:<br>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.)<br>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.)</td></tr>
<tr><td colspan="2">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.<br><br>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.</td></tr>
<tr><td colspan="2">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?</td></tr>
</table>

| |
|---|
| Not in scope. 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.<br>The records will not, as part of this experiment, be made public. |

| | |
|---|---|
| h) Copyright, licensing, rights, and/or privacy restrictions<br>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.<br>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. |
| Where possible, we will use cross-evaluation when training models on LoC data in order to avoid introducing selection bias or overfitting the model to the training set. If this is not possible, the dataset will be explicitly split into training, validation and test data without cross-evaluation. The split will be random, and follow a standard 80/10/10 split. We would expect the training, validation, and test_data to comprise examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset.<br>However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.<br>In the case of the initial evaluation of the pre-trained GROBID models we will evaluate against the entire set of ebooks. |
| b) For training data:<br>1) if the model is pre-trained, describe the data on which it was trained; |

2) if the model will be fine-tuned, outline the data involved in this process;

3) if the model is being trained from scratch, outline the plan for creating training data.

We would expect to:

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

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

The models that are distributed with GROBID are trained on a relatively small, manually labeled dataset that has been selected for "accuracy and coverage", aiming to cover the wide range of edge-cases discovered during development. This includes documents from domains and publishers that are otherwise under-represented in existing datasets (consisting of preprints, medical journals), v. https://grobid.readthedocs.io/en/latest/Principles/#training-data-qualitat-statt-quantitat

The training corpus for GROBID is different from LoC ebooks in terms of broad subject coverage and document structure, so we would expect to see improvements pre- and post- training on LoC ebooks. Validating whether this is the case or not will be one of the project outputs.

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. |
|---|
| For this experiment, the goal is to test, in a time-limited period, the success of these models in matching existing human catalogers at generating bibliographic metadata from ebooks. The type of task being carried out in this experiment is less likely to surface bias, as we are primarily looking for *existing* text in an existing record, and will be fine-tuning models based on existing catalog records.<br>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 | GROBID (GeneRation Of BIbliographic Data) - a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications. The extraction includes bibliographical information (e.g. title, abstract, authors, affiliations, keywords) along with the text and document structure. |
| b) Intended use | 1) Automated extraction of bibliographic metadata from ebooks.<br>2) As a means of extracting textual and layout data from ebooks. |
| c) Limitations | GROBID has been developed to target scientific and technical publications, and evaluation has been carried out against datasets of this type.<br>The models that are distributed with GROBID are trained on a relatively small dataset that has been selected for "accuracy and coverage", aiming to cover the wide range of edge-cases discovered during development. This includes documents from domains and publishers that |

| | |
|---|---|
| | are otherwise under-represented in existing datasets (preprints, medical journals).<br>Cannot process EPUB format files, making conversion to PDF necessary for files of that type. |
| d) Copyright and licensing details for the model | GROBID is released under the Apache License 2.0<br>https://github.com/kermitt2/grobid/blob/master/LICENSE |
| e) Link to documentation | Documentation: https://grobid.readthedocs.io/en/latest/<br>Github: https://github.com/kermitt2/grobid |
| 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.<br>From the data we will be gathering standard metrics as part of the process, including:<br>● Precision<br>● Recall<br>● F1-Score<br>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 at the end of evaluation of the performance of the models shipped with GROBID on a test set of the LoC ebook dataset. Further metrics will be gathered during a training and evaluation of GROBID trained on the LoC ebook dataset. |
| 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. However, for extreme multilabel classification and token classification of the type being tested in this experiment, this is probably an unreasonably high threshold.

It is important to note here, that with regards to subject indexing in particular, inter-indexer consistency for human indexers is often empirically measured at significantly under 100%. Figures of 30-50% are sometimes cited. On this measure, an automated subject indexing tool that matched inter indexer consistency for human indexers might be considered "good enough".

See: https://researchcommons.waikato.ac.nz/handle/10289/3513 for example, for a PhD thesis comparing consistency between human and machine indexing,  https://eric.ed.gov/?id=EJ956121 on inter-indexer consistency in subject cataloging, or https://eric.ed.gov/?id=ED413903 which also includes non-subject cataloging such as Marc 245 (Title) fields.

There are also trade-offs. For example, it is possible to set the threshold for membership in some classes higher, which will improve the overall *precision* of the pipeline while potentially reducing the *recall* of the pipeline.

Doing this will reduce the number of false positives: fewer ebooks falsely identified as a subject category that does not apply. However, this will be at the cost of increasing the number of false negatives: fewer ebooks will be correctly identified as that subject category than otherwise would have been. If we increase the threshold for being classified as "Post-communism" for example, fewer documents will be *incorrectly* classified as having the subject "Post-communism" at the cost of missing some documents that *should* have been classified as "Post-communism".

Choosing where to set the threshold is partly a matter of institutional policy. The LoC might decide to be quite permissive, for example, with subject tagging to drive discovery, or may decide to be more cautious in order to avoid irrelevant results.

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.

For the initial evaluation of the pre-trained GROBID models we will:

1. Take the PDFs stored in an Amazon S3 bucket. (v. Section B(e) above for epub conversion to PDF)
2. Process each with GROBID's pre-trained models, producing a TEI XML document containing bibliographic metadata extracted from the PDF.
3. Extract resulting field data from the TEI XML document and use to generate metrics (F-Score, Precision etc.) for each primary field being targeted (n.b. these being the set of bibliographic metadata fields provided by the pre-trained models).
4. Store the generated catalog data and metrics for use in the final report.

For the training of the GROBID "bibliographic" models we will:

1. Use cross-evaluation to successively train GROBID on the full corpus of PDFs. In the event that this proves problematic, or looks like it is likely to overfit the model, we will take a *Training* set of PDFs and MARCXML data stored in an Amazon S3 bucket.
2. Create training data from those PDFs using GROBID utilities, including the targeted metadata fields from the MARCXML.
3. Run a training cycle for each of the targeted "bibliographic" models using this training data.
4. Store the resulting model.

Once these models have been created we will:

1. Run the  workflow across the *Test* set of ebooks (or the complete set if we are doing cross-evaluation) 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.
2. Generate metrics (F-Score, Precision, etc) for each of the primary field types in the records.
3. 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.
- An instance of Digirati's Django-based "Task Service" for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML. This "Task Service" will also run the final composite workflow across the ebook test set.

# 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:

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

This particular data processing plan concerns the testing of machine learning models (or other computational approaches) for generating cataloging metadata from Library ebooks.

The goal is, for this model, to:

- measure the quality of the outputs (using some standard metrics)
- measure the cost (in terms of hours of person time, and in terms of compute costs)
- 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

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 hyperparameters 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)

- 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)

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.

While the goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible, with the minimum fields to be generated comprising:

- Titles
- Author names,
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form
- Subject terms.

It is not always the case, however, that each individual ML model is going to be suited to all of these.

Models that are suited to, for example, extreme multi-label text classification—such as generating subject or genre terms—are not necessarily the same models that are suited to named entity extraction or token classification, such as identifying an ISBN, copyright statement, or Title for an ebook. While some models under test may attempt to generate all of the above types of catalog metadata, in some cases they will not.

In the case of Annif, the toolkit is intended for the extraction of subject indexing and classification metadata from input documents, with support for different subject indexing algorithms and vocabularies. The expected range of fields that Annif models will generate is:

- Genre/form (using LCFGT as the vocabulary / thesaurus)
- Subject terms (using LCSH as the vocabulary / thesaurus)

We will not be generating other bibliographic metadata as part of this experiment.

The broad aim for this experiment is to assess the performance of Annif using different backends when trained with the LoC ebook dataset and the associated subject terms. As this is targeting only a few fields, evaluation will be carried out on these fields rather than all metadata fields.

**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:

- Generated subject and genre terms for each ebook, with their accompanying identifiers in LCSH or 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 subject terms to the actual catalog records for the same ebook (see attachment below). These will be provided as JSON files, and as CSV/XLSX files.

**A4: Description of intended use**

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

The experiment will evaluate performance of Annif when trained and tested with plain text extracted from the LoC ebook dataset, with the resulting models being used to generate subject terms for the ebooks.

The primary intended use for the data generated is as part of the final report, rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains.<br>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 | The dataset consists of ebooks and MarcXML files with catalog records for those ebooks.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs<br>c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records |

| | |
|---|---|
| or it may be a series of folders containing images derived from a particular source. | iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br><br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br><br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance<br>1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | The information in this dataset originated from four collections of LoC ebooks:<br><br>1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br><br>Further details to be provided by LoC. |
| d) Compilation methods<br>1. How is/was this dataset compiled, when, and by whom?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br><br>Further details to be provided by LoC. |
| e) Preprocessing steps | 1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were: |

| | |
|---|---|
| 1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>3. Is the data organized according to a schema, content standard, or other standard? If yes, which one? | a. Conversion of PDF and epub to plaintext using a mixture of tools, including Grobid (see other data processing plan) and Pandoc.<br>b. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any)<br>c. Normalize whitespace<br><br>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)..<br><br>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?

Not in scope. 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.

| | |
|---|---|
| h) Copyright, licensing, rights, and/or privacy restrictions<br>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.<br>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. |
| Where possible, we will use cross-evaluation when training models on LoC data in order to avoid introducing selection bias or overfitting the model to the training set. If this is not possible, the dataset will be explicitly split into training, validation and test data without cross-evaluation. The split will be random, and follow a standard 80/10/10 split. We would expect the training, validation, and test_data to comprise examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset.<br>However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test. |
| b) For training data:<br>1) if the model is pre-trained, describe the data on which it was trained;<br>2) if the model will be fine-tuned, outline the data involved in this process;<br>3) if the model is being trained from scratch, outline the plan for creating training data. |
| We would expect to:<br>1. Train the model(s) based on the training subset of the LoC ebook dataset<br>2. Testing the LoC-trained models on a test subset of the LoC ebook dataset.<br>3. Produce scores/metrics for each record, and for the collection in aggregate for each testing cycle.<br>Each of the training and fine-tuning steps will use the text from the books and the MarcXML records. |
| 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.<br>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. |
| For this experiment, the goal is to test, in a time-limited period, the success of these models in matching existing human catalogers at generating bibliographic metadata from ebooks. The type of task being carried out in this experiment is less likely to surface bias, as we are primarily looking for *existing* text in an existing record, and will be fine-tuning models based on existing catalog records.<br>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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | Annif: automated subject indexing toolkit |
| b) Intended use | Automated creation of subject/genre information for texts. |
| c) Limitations | Annif will only allow us to target metadata fields which are good candidates for the kind of multi-label text classification provided by the toolkit.<br>Annif provides a number of different backends for multi-label text classification, and allows for combinations of different backends via ensembles. This presents a wide range of |

| | |
|---|---|
| | possible usage, and time-boxed experiments with these models/approaches may not exhaust all of the possibilities. |
| d) Copyright and licensing details for the model | Annif is released under the Creative Commons Attribution 4.0 International License<br>https://github.com/NatLibFi/Annif-tutorial/blob/master/LICENSE.txt |
| e) Link to documentation | Documentation: https://github.com/NatLibFi/Annif/wiki<br>Github: https://github.com/NatLibFi/Annif<br>Article: https://www.jlis.it/index.php/jlis/article/view/437 |
| 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.<br>From the data we will be gathering standard metrics as part of the process, including:<br>● Precision<br>● Recall<br>● F1-Score<br>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.<br>The Annif model(s) for this experiment will be fine-tuned on data, hyperparameters tuned, and then the model will run once over the test set and final metrics will be generated. |
| 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. However, for extreme multilabel classification and token classification of the type being tested in this experiment, this is probably an unreasonably high threshold.

It is important to note here, that with regards to subject indexing in particular, inter-indexer consistency for human indexers is often empirically measured at significantly under 100%. Figures of 30-50% are sometimes cited. On this measure, an automated subject indexing tool that matched inter indexer consistency for human indexers might be considered “good enough”.

See: https://researchcommons.waikato.ac.nz/handle/10289/3513 for example, for a PhD thesis comparing consistency between human and machine indexing, https://eric.ed.gov/?id=EJ956121 on inter-indexer consistency in subject cataloging, or https://eric.ed.gov/?id=ED413903 which also includes non-subject cataloging such as Marc 245 (Title) fields.

There are also trade-offs. For example, it is possible to set the threshold for membership in some classes higher, which will improve the overall *precision* of the pipeline while potentially reducing the *recall* of the pipeline.

Doing this will reduce the number of false positives: fewer ebooks falsely identified as a subject category that does not apply. However, this will be at the cost of increasing the number of false negatives: fewer ebooks will be correctly identified as that subject category than otherwise would have been. If we increase the threshold for being classified as “Post-communism” for example, fewer documents will be *incorrectly* classified as having the subject “Post-communism” at the cost of missing some documents that *should* have been classified as “Post-communism”.

Choosing where to set the threshold is partly a matter of institutional policy. The LoC might decide to be quite permissive, for example, with subject tagging to drive discovery, or may decide to be more cautious in order to avoid irrelevant results.

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.

For a single iteration of training and evaluation of a Annif model(s) we test we will:

1. Create a Training dataset consisting of:
    a. Subject vocabulary data from LCFGT/LCSH
    b. A training corpus of plaintext extracted from pdf/epub annotated with subject and genre terms from the MARCXML.
2. Create project configurations for the pipeline/model
3. Run the training cycle
4. Fine-tune hyperparameters
5. Store the resulting model
6. Run the  workflow across the *Test* set of ebooks (or the entire corpus if we are doing cross-evaluation) generating subject terms 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.
7. Generate metrics (F-Score, Precision, etc) for each of the targeted field types in the records.
8. Store the generated catalog data and metrics for use in the final report

Note, that Annif can use SKOS vocabularies as part of the workflow, and can use, in some of the models, the relationships between tags within the vocabulary, which means that it is possible that some “incorrect” subject tags may be instances of a broader or narrower term within the LCSH vocabulary, and would not be judged to be incorrect by end users or a cataloger. As a stretch goal, we may check whether “incorrect” subject tags are related terms (broader/narrower, etc) and gather some additional metrics to quantify this.

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.
- An instance of Digirati’s Django-based “Task Service” for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML. This “Task Service” will also run the final composite workflow across the ebook test set.

# 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:

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

This particular data processing plan concerns the testing of machine learning models (or other computational approaches) for generating cataloging metadata from Library ebooks.

The goal is, for this model, to:

- measure the quality of the outputs (using some standard metrics)
- measure the cost (in terms of hours of person time, and in terms of compute costs)
- 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

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 hyperparameters 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)

- 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)

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.

While the goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible, with the minimum fields to be generated comprising:

- Titles
- Author names,
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form
- Subject terms.

It is not always the case, however, that each individual ML model is going to be suited to all of these.

Models that are suited to, for example, extreme multi-label text classification—such as generating subject or genre terms—are not necessarily the same models that are suited to named entity extraction or token classification, such as identifying an ISBN, copyright statement,  or Title for an ebook. While some models under test may attempt to generate all of the above types of catalog metadata, in some cases they will not.

In the case of this particular model, “LoC Spacy 3”, the expected scope is that the model will generate:

- Titles
- Author Names
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form (using LCFGT as the vocabulary / thesaurus)
- Subject terms (using LCSH as the vocabulary / thesaurus)

For each ebook, this model “LoC Spacy 3” will provide a list of suggested subject or genre terms, along with the relevant identifier for that subject term, where available. The model will also attempt to extract the spans (sequences of words) that match the bibliographic metadata fields (Title, Author Names, etc) and tag them with their category/field label.

N.B. While we are using a single library/framework for this experiment, different fields will be handled by different pipeline components.

**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 subject and genre terms for each ebook, with their accompanying identifiers in LCSH or LCGFT.
- N.B. The interim data format will not be Marc but a simplified JSON representation that we can convert into Marc later.
- A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file.
  - In the case of this experiment, which uses Spacy as the NLP engine, these will take the form of config (*.cfg*) files in the Confection format used by Spacy along with any other project files, such as the Project.yml YAML file used to define the project.
- Exports of the data models generated (where possible).
  - In the case of this experiment, which uses Spacy to train NLP models the export will be in the form produced by the helper methods provided by Spacy's Project architecture and using Spacy's *.to_disk* serializer methods for the pipeline steps and models. The primary goal here is to capture and preserve the state of the model and pipelines at the end of the experiment, rather than to package the model for production use.
- 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.

**A4: Description of intended use**

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

The experiment will train models on the ebook plaintext and the resulting models will be used to generate the other data. The models generated will be tested and refined—tuning hyperparameters, for example—and then used to generate catalog records.

The primary intended use for the data generated is as part of the final report, rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains.<br>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.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs<br>c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance<br>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:<br>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<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br><br>Further details to be provided by LoC. |
| d) Compilation methods<br>1. How is/was this dataset compiled, when, and by whom?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br><br>Further details to be provided by LoC. |
| e) Preprocessing steps<br>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>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:<br>a. Conversion of PDF and epub to plaintext using a mixture of tools, including Grobid (see other data processing plan) and Pandoc.<br>b. Normalization of character set encoding (conversion to unicode for files that aren’t encoded as unicode, if any)<br>c. Normalize whitespace<br><br>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 pipeline stages in the Spacy pipeline.<br><br>For example, we actively do not want to remove stop words or lemmatize the text if our goal is to extract the Title from the text. We *will* want to remove stop words and lemmatize the text, if we are generating subject or genre information.<br><br>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)..<br>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?

Not in scope. 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.

| | |
|---|---|
| h) Copyright, licensing, rights, and/or privacy restrictions<br>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.<br>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.

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

We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. We may also split out digitized from born-digital material at the test stage, in order to have comparative data.

However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.

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

This experiment uses Spacy, a widely used NLP library. Spacy can use multiple language models, and we would expect to:

1. Evaluate several candidate base language models as part of the experiment. We will use the large English model and the RoBERTa based Spacy transformer model.
2. Train the model(s) based on the training subset of the data
3. Further fine-tune the model(s) based on the validation set of data
4. Test the model on the test set and produce scores/metrics for each record, and for the collection in aggregate.

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

Regarding pre-training, Spacy provides a list of the language models here: https://spacy.io/models/en

The core English models are pre-trained on OntoNotes 5, ClearNLP Constituent-to-Dependency Conversion (Emory University), WordNet 3.0 (Princeton University) and Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl) (Explosion) datasets. We would expect to use the large English language model, and the transformer based model for English records.

In addition, the Spacy transformer based language model is based on RoBERTa, so is additionally trained on the RoBERTa base dataset (see also https://huggingface.co/roberta-base).

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.<br>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. |
| N/A |

## 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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | LoC Spacy: Spacy (spaCy · Industrial-strength Natural Language Processing in Python) with additional pipeline steps for LoC catalog metadata. |
| b) Intended use | Automated extraction of bibliographic metadata and subject/genre classifications from ebooks. |
| c) Limitations | Spacy is primarily a natural language processing library, so the primary input is plaintext only.<br>Visual information (size, font style, location on page, location within the book structure, etc) present in the ebooks is out of scope for this experiment (although will be tested in a different experiment). |

| | |
|---|---|
| | Some Spacy pipeline components benefit from smaller blocks of text, so we may need to select a subset of the ebooks (the first N thousand words, for example), or process the books in chunks. |
| d) Copyright and licensing details for the model | Spacy is licensed under the MIT license, see: https://github.com/explosion/spaCy/blob/master/LICENSE |
| e) Link to documentation | https://spacy.io/ |
| 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.<br>From the data we will be gathering standard metrics as part of the process, including:<br>● Precision<br>● Recall<br>● F1-Score<br>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.<br>The Spacy model(s) for this experiment will be trained on data, iterated over and fine-tuned—metrics will be gathered at this point but they are part of an automated finetuning process and will not be persisted— and then the model will run once over the test set and final metrics will be generated. |
| 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. However, for extreme multilabel classification and token classification of the type being tested in this experiment, this is probably an unreasonably high threshold.

It is important to note here, that with regards to subject indexing in particular, inter-indexer consistency for human indexers is often empirically measured at significantly under 100%. Figures of 30-50% are sometimes cited. On this measure, an automated subject indexing tool that matched inter indexer consistency for human indexers might be considered "good enough".

See: https://researchcommons.waikato.ac.nz/handle/10289/3513 for example, for a PhD thesis comparing consistency between human and machine indexing, https://eric.ed.gov/?id=EJ956121 on inter-indexer consistency in subject cataloging, or https://eric.ed.gov/?id=ED413903 which also includes non-subject cataloging such as Marc 245 (Title) fields.

There are also trade-offs. For example, it is possible to set the threshold for membership in some classes higher, which will improve the overall *precision* of the pipeline while potentially reducing the *recall* of the pipeline. Doing this will reduce the number of false positives: fewer ebooks falsely identified as a subject category that does not apply. However, this will be at the cost of increasing the number of false negatives: fewer ebooks will be correctly identified as that subject category than otherwise would have been. If we increase the threshold for being classified as "Post-communism" for example, fewer documents will be *incorrectly* classified as having the subject "Post-communism" at the cost of missing some documents that *should* have been classified as "Post-communism".

Choosing where to set the threshold is partly a matter of institutional policy. The LoC might decide to be quite permissive, for example, with subject tagging to drive discovery, or may decide to be more cautious in order to avoid irrelevant results.

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.

Spacy provides multiple APIs for different pipeline components which can be trained individually and/or combined to provide different kinds of data.

We propose to employ four different Spacy pipeline components to extract the data from the ebook plaintext, with each component trained on data taken from the MarcXML records.

These are as follows:

- Entity/Span Ruler: For certain kinds of fields which have regular and predictable forms such as LCCNs or ISBNs, or other identifiers, we will write and test rule-based approaches for identifying these fields in the ebook text.
- Span Categorizer: For other fields such as Author names, Titles, Publisher names, dates, rights statements, etc we will use Spacy's Span Categorizer component to *train* a new pipeline component using the Marc records and the ebook texts.
- Named Entity Recognizer: A similar component, the Entity Recognizer can also be trained, and we would propose to use this approach, too, and compare the outputs. Our expectation is that this component will work well for certain kinds of data: dates, and people's names, but may work less well for longer and less discrete fields such as Titles or rights statements.
- Text Categorizer: For fields that are a property of the entire document, such as Subjects and Genres, we propose to train a TextCategorizer model to identify LCSH subjects and LCGFT genres from the document text.

Spacy provides a project structure and format for:

- Creating and storing a corpus of texts (for training, evaluation, and test)
- Defining the order of processing steps, including text cleaning, pre-processing, etc.
- Training and fine tuning models
- Exporting models for reuse

We will:

1. Take plaintext extracted from PDFs and ePubs stored in an Amazon S3 bucket
2. Create Spacy corpora from these (as Spacy *DocBin* files) also stored in an Amazon S3 bucket for reuse
3. Create project configurations for each of the pipelines above
4. Run the training cycle
5. Run the optimization cycle / eval cycle (where Spacy refines the model outputs and tunes parameters)
6. Store the resulting model
7. Store Spacy's metric data (output as part of the model evaluation)

After each of the models is complete and tested we will:

8. Create a single Spacy workflow with pipeline steps for each of the above to create a "automatic cataloging" workflow pipeline for all of the core Marc fields in scope for the Task Order. Taking, for each field, the highest scoring pipeline component if more than one component can produce the same data.

9. 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.
10. Generate metrics (F-Score, Precision, etc) for each of the primary field types in the records.
11. 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, Spacy DocBin data files (for corpora), Spacy models
- An instance of Digirati's Django-based "Task Service" for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML.

Note:

1. There is no annotation involved in this experiment
2. The project consists of a mixture of supervised learning:
    a. Span Categorizer
    b. Entity Recognizer
    c. Text Categorizer
3. And non-trained/rule-based approaches (SpanRuler, EntityRuler) for identifiers and other regular fields.

# 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:

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

This particular data processing plan concerns the testing of machine learning models (or other computational approaches) for generating cataloging metadata from Library ebooks.

The goal is, for this model, to:

- measure the quality of the outputs (using some standard metrics)
- measure the cost (in terms of hours of person time, and in terms of compute costs)
- 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

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 hyperparameters 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)

- 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)

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.

While the goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible, with the minimum fields to be generated comprising:

- Titles
- Author names,
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form
- Subject terms.

It is not always the case, however, that each individual ML model is going to be suited to all of these.

Models that are suited to, for example, extreme multi-label text classification—such as generating subject or genre terms—are not necessarily the same models that are suited to named entity extraction or token classification, such as identifying an ISBN, copyright statement,  or Title for an ebook. While some models under test may attempt to generate all of the above types of catalog metadata, in some cases they will not.

In the case of this particular model,  BERT, the expected scope is that the model will generate:

- Titles
- Author Names
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form (using LCFGT as the vocabulary / thesaurus)
- Subject terms (using LCSH as the vocabulary / thesaurus)

BERT is a transformer based NLP model developed by Google. By BERT we are referring to a family of models, including DistilBERT, RoBERTa, etc. and in the document that follows, we will use BERT as a shorthand for all of these. The broad aim for this experiment is to assess transformer based approaches to text and token classification on LoC ebooks.

BERT, on its own, will not provide any of the above bibliographic metadata fields out-of-the-box. Instead, we will leverage BERT following standard practice for using transformer based models for:

1. Entity Recognition/Token Classification: Multiple examples in the literature of BERT used for token/span classification exist for labeling text with entity types (author, date, title, etc).
2. Topic modeling: Similarly, multiple examples of BERT used for text classification exist in the literature. BERTopic, for example, provides a range of methods for training and evaluating models for topic (i.e. subject and genre) classification.

For each ebook, this model, LoC BERT, will provide a list of suggested subject or genre terms, along with the relevant identifier for that subject term, where available. The model will also attempt to extract the spans (sequences of words) that match the bibliographic metadata fields (Title, Author Names, etc) and tag them with their category/field label.

**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 subject and genre terms for each ebook, with their accompanying identifiers in LCSH or LCGFT.
- N.B. The interim data format will not be Marc but a simplified JSON representation that we can convert into Marc later.
- 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). BERTopic, for example, provides standard methods for saving models to disk.
- 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.

**A4: Description of intended use**

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

The experiment will train models on the ebook plaintext and the resulting models will be used to generate the other data. The models generated will be tested and refined—tuning hyperparameters, for example—and then used to generate catalog records.

The primary intended use for the data generated is as part of the final report, rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains.<br>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.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs<br>c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance | The information in this dataset originated from four collections of LoC ebooks: |

| | |
|---|---|
| 1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | 1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br><br>Further details to be provided by LoC. |
| d) Compilation methods<br>1. How is/was this dataset compiled, when, and by whom?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br><br>Further details to be provided by LoC. |
| e) Preprocessing steps<br>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>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:<br>a. Conversion of PDF and epub to plaintext using a mixture of tools, including Grobid (see other data processing plan) and Pandoc.<br>b. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any)<br>c. Normalize whitespace<br><br>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.<br><br>For example, we actively do not want to remove stop words or lemmatize the text if our goal is to extract the Title from the text. We *will* want to remove stop words and lemmatize the text, if we are generating subject or genre information using BERTopic, for example. |

<table>
<tr><td></td><td>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)..<br><br>3. The metadata is organized as MarcXML files following usual LoC cataloging practice.</td></tr>
<tr><td colspan="2">f) Potential risks to people, communities and organizations & strategies for risk mitigation:<br>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.)<br>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.)</td></tr>
<tr><td colspan="2">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.<br><br>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.</td></tr>
<tr><td colspan="2">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?</td></tr>
<tr><td colspan="2">Not in scope. 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.<br><br>The records will not, as part of this experiment, be made public.</td></tr>
<tr><td>h) Copyright, licensing, rights, and/or privacy restrictions<br>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.</td><td>The material comprises a mixture of open access and copyrighted ebooks.<br><br>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.</td></tr>
</table>

*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.

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

We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. We may also split out digitized from born-digital material at the test stage, in order to have comparative data.

However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.

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

We would expect to:

1. Evaluate several candidate transformer based large language models as part of the experiment, including BERT, RoBERTa, distilBERT, etc.
2. Train the model(s) based on the training subset of the data
3. Further fine-tune the model(s) based on the validation set of data
4. Test the model on the test set and produce scores/metrics for each record, and for the collection in aggregate.

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

The BERT model was pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers). Regarding pre-training, RoBERTa, is additionally trained on the RoBERTa base dataset (see also https://huggingface.co/roberta-base). DistilBERT is trained on the same dataset but the size of the model is reduced.

BERT based models trained for named entity recognition have been fine-tuned on the CoNLL-2003 Named Entity Recognition dataset, and we will further fine-tune these on the LoC dataset (plaintext, and MarcXL 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.

BERT models have known biases that originate in the datasets they were trained on. See: https://dmccreary.medium.com/showing-bias-in-bert-475e98dabf51 on gender bias in BERT, and https://towardsdatascience.com/racial-bias-in-bert-c1c77da6b25a on race bias in BERT.

Note, that this bias will tend to show up in masking tasks or text completion tasks where the language model is being asked to generate text.

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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | BERT. |

| | |
|---|---|
| b) Intended use | Automated extraction of bibliographic metadata and subject/genre classifications from ebooks. |
| c) Limitations | BERT models are trained on a wide range of source data: BookCorpus, Wikipedia, CONLL, etc. Large language models of this type are very powerful, but require fine-tuning for specific use cases. A time-boxed experiment with these models/approaches may not exhaust all of the possibilities.<br>BERT models typically have a maximum text size they can work with, so we will need to preprocess the data to identify good quality short text extracts to train the models on. There are approaches that can be adopted to partly mitigate against BERTs text length limits. |
| d) Copyright and licensing details for the model | BERT and derivative models are generally released under open source licenses. https://github.com/google-research/bert/blob/master/LICENSE |
| e) Link to documentation | Multiple sources, including: https://github.com/google-research/bert and https://huggingface.co/docs/transformers/model_doc/distilbert and https://huggingface.co/docs/transformers/model_doc/roberta |
| 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.<br>From the data we will be gathering standard metrics as part of the process, including:<br>● Precision<br>● Recall<br>● F1-Score<br>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 |

<table>
<tr><td>h) Audit schedule (how often and how many times will performance metrics be checked?)</td><td>We would expect to gather metrics once at the end of the training and evaluation cycle.<br>The BERT model(s) for this experiment will be fine-tuned on data, hyperparameters tuned, and then the model will run once over the test set and final metrics will be generated.</td></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td colspan="2">Part of this experiment is to assist the LoC to determine what counts as successful algorithmic performance.<br>For example, general figures such as 70% accuracy are sometimes used to define “ideal” or “good” model performance. However, for extreme multilabel classification and token classification of the type being tested in this experiment, this is probably an unreasonably high threshold.<br>It is important to note here, that with regards to subject indexing in particular, inter-indexer consistency for human indexers is often empirically measured at significantly under 100%. Figures of 30-50% are sometimes cited. On this measure, an automated subject indexing tool that matched inter indexer consistency for human indexers might be considered “good enough”.<br>See: https://researchcommons.waikato.ac.nz/handle/10289/3513 for example, for a PhD thesis comparing consistency between human and machine indexing,  https://eric.ed.gov/?id=EJ956121 on inter-indexer consistency in subject cataloging, or https://eric.ed.gov/?id=ED413903 which also includes non-subject cataloging such as Marc 245 (Title) fields.<br>There are also trade-offs. For example, it is possible to set the threshold for membership in some classes higher, which will improve the overall <i>precision</i> of the pipeline while potentially reducing the <i>recall</i> of the pipeline.<br>Doing this will reduce the number of false positives: fewer ebooks falsely identified as a subject category that does not apply. However, this will be at the cost of increasing the number of false negatives: fewer ebooks will be correctly identified as that subject category than otherwise would have been. If we increase the threshold for being classified as “Post-communism” for example, fewer documents will be <i>incorrectly</i> classified as having the subject “Post-communism” at the cost of missing some documents that <i>should</i> have been classified as “Post-communism”.<br>Choosing where to set the threshold is partly a matter of institutional policy. The LoC might decide to be quite permissive, for example, with subject tagging to drive discovery, or may decide to be more cautious in order to avoid irrelevant results.<br>Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report.</td></tr>
<tr><td colspan="2">i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning.</td></tr>
</table>

We will, for each large language model we test (BERT, distilbert, RoBERTa, etc) 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 (BERT models typically have a 500 token max length)
    b. Tag these sequences with their label (Title, ISBN, LCCN, etc)
    c. Format these into the right corpus/training format (for example, as JSONLines data files that we can easily parse, or as more compressed tabular data formats for persistence in S3)
3. Create project configurations for each of the pipelines above
4. Run the training cycle
5. Fine-tune hyperparameters
6. Store the resulting model

We would expect to use BERTopic for subject and genre terms only. BERTopic can work on the entire text of a document as it uses keywording techniques to characterize a document so we won't have to work with a 500 token limit.

See also: https://huggingface.co/BritishLibraryLabs/bl-books-genre for an example of how the British Library Labs team have used distilBERT to identify book genres by training on book titles.

For token/span classification (for identifiers, Titles, etc) will use custom code to fine-tune a transformer based large language model for LoC metadata. This will make use of the transformer pipelines already provided by the APIs:

1. Named Entity Recognition Pipelines
2. Text Classification Pipelines

These models will require us to process smaller segments of text, or  adopt some techniques to chunk or preprocess the data and then reintegrate the outputs later. See this Medium.com article on chunking for transformers or using something like AllenAI's Longformer approach to transforming models for longer texts.

After each of the models is complete and tested we will:

7. 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.  N.B. These are not likely to comprise a single pipeline implemented using a single model and approach. Rather, we will create custom scripts that process the data sequentially, as appropriate.
    a. For example, run the entire text through a trained BERTopic model to generate Subject and Genre fields.
    b. Segment the text into smaller blocks (per line, or per paragraph) or use Longformer or other similar approaches and run them through one or more trained models for:
        i. Title, Author name, ISBN, LCCN, etc

c. Combine the data from the project pipeline steps into a single catalog record

d. N.B. These scripts will be run using the Django based “Task Service” we will deploy as part of the project.

8. 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.
9. Generate metrics (F-Score, Precision, etc) for each of the primary field types in the records.
10. 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.
- An instance of Digirati’s Django-based “Task Service” for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML. This “Task Service” will also run the final composite workflow across the ebook test set.

# 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:

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

This particular data processing plan concerns the testing of machine learning models (or other computational approaches) for generating cataloging metadata from Library ebooks.

The goal is, for this model, to:

- measure the quality of the outputs (using some standard metrics)
- measure the cost (in terms of hours of person time, and in terms of compute costs)
- 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

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 hyperparameters 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)

- 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)

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.

While the goal of the set of experiments as a whole, is to generate full level bibliographic records whenever possible, with the minimum fields to be generated comprising:

- Titles
- Author names,
- Unique identifiers
- Date of issuance
- Date of creation
- Genre/form
- Subject terms.

It is not always the case, however, that each individual ML model is going to be suited to all of these.

Models that are suited to, for example, extreme multi-label text classification—such as generating subject or genre terms—are not necessarily the same models that are suited to named entity extraction or token classification, such as identifying an ISBN, copyright statement,  or Title for an ebook. While some models under test may attempt to generate all of the above types of catalog metadata, in some cases they will not.

Building on previous experiments using NLP, this experiment will include features derived from layout data obtained from processing the ebooks with GROBID or similar tools such as pdfalto (which generates ALTO XML files from PDF). This will allow data that contains semantic information about the text based on text size, location on page, location within the book structure etc. to be used as part of the classification of text spans and the identification of metadata about the ebook.

The expected scope is that the model will generate:

- Titles
- Author Names
- Unique identifiers
- Date of issuance
- Date of creation

Fields such as subject or genre, which are not properties of specific spans of text within the document are not in scope.

**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.
- N.B. The interim data format will not be Marc but a simplified JSON representation that we can convert into Marc later.
- A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file, in a format dependent on the NLP library used.
- Exports of the data models generated in a format dependent on the NLP library used.
- 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.

**A4: Description of intended use**

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

The experiment will train models on the ebook pdfs and epubs and the resulting models will be used to generate the other data. The models generated will be tested and refined—tuning hyperparameters, for example—and then used to generate catalog records.

The primary intended use for the data generated is as part of the final report, rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive | The dataset consists of ebooks and MarcXML files with catalog records for those ebooks.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs |

| | |
|---|---|
| information about what intellectual content the dataset contains.<br>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. | c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance<br>1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | The information in this dataset originated from four collections of LoC ebooks:<br>1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br>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?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br>Further details to be provided by LoC. |
| e) Preprocessing steps<br>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>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:<br>a. Conversion of epub to PDF using e.g. Pandoc or Calibre.<br>b. Conversion of PDF files to TEI XML using GROBID, the structured output of which will give us data on the layout of the text.<br>c. Normalization of character set encoding (conversion to unicode for files that aren't encoded as unicode, if any)<br>d. Normalize whitespace<br>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)..<br>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:<br>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.)<br>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.<br>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? |
| Not in scope. 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.<br>The records will not, as part of this experiment, be made public. |

| | |
|---|---|
| h) Copyright, licensing, rights, and/or privacy restrictions<br>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.<br>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. |

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

We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. We may also split out digitized from born-digital material at the test stage, in order to have comparative data.

However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.

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

This experiment uses Spacy or BERT, and  is intended to supplement the token/span categorisation available from them with additional layout data.

1. Evaluate several candidate base language models as part of the experiment. We will use the large English model and the RoBERTa based Spacy transformer model.
2. Train the model(s) based on the training subset of the data
3. Further fine-tune the model(s) based on the validation set of data
4. Test the model on the test set and produce scores/metrics for each record, and for the collection in aggregate.

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

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.

N/A

## 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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | NLP With Layout: Custom pipeline steps for best-candidate NLP model identified as part of other experiments (see data processing plans for Spacy and BERT).<br>These pipeline steps will include layout features as part of text tokenization/parsing and make use of them as part of token/span classification. |
| b) Intended use | Automated extraction of bibliographic metadata from ebooks. |
| c) Limitations | Limited in the layout features that are available as output from GROBID, namely bounding boxes for text and other features on a PDF page and position of the text within the typeset document as a whole. For born-digital content of the type found in the LoC ebooks dataset these features do carry semantic value, and allow for testing of the utilization of layout data without the overhead of treating the PDF content as visual data.<br>There is not a dataset of documents consisting of text and layout information annotated with features matching the targeted metadata fields.<br>Due to the likelihood that notable bibliographic metadata, identified by layout and related to the ebook itself is likely to occur earlier in the text, this approach may be better targeted at a |

| | |
|---|---|
| | limited number of pages from the start of ebooks.<br>Will be used as a supplement to a primarily natural language processing model. |
| d) Copyright and licensing details for the model | Will use models/libraries/frameworks released under a FOSS license. |
| e) Link to documentation | N/A |
| 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.<br>From the data we will be gathering standard metrics as part of the process, including:<br>● Precision<br>● Recall<br>● F1-Score<br>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.<br>The model(s) for this experiment will be trained on data, iterated over and fine-tuned—metrics will be gathered at this point but they are part of an automated finetuning process and will not be persisted— and then the model will run once over the test set and final metrics will be generated. |
| 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. However, for extreme multilabel classification and token classification of the type being tested in this experiment, this is probably an unreasonably high threshold.

It is important to note here, that with regards to subject indexing in particular, inter-indexer consistency for human indexers is often empirically measured at significantly under 100%. Figures of 30-50% are sometimes cited. On this measure, an automated subject indexing tool that matched inter indexer consistency for human indexers might be considered “good enough”.

See: https://researchcommons.waikato.ac.nz/handle/10289/3513 for example, for a PhD thesis comparing consistency between human and machine indexing, https://eric.ed.gov/?id=EJ956121 on inter-indexer consistency in subject cataloging, or https://eric.ed.gov/?id=ED413903 which also includes non-subject cataloging such as Marc 245 (Title) fields.

There are also trade-offs. For example, it is possible to set the threshold for membership in some classes higher, which will improve the overall *precision* of the pipeline while potentially reducing the *recall* of the pipeline. Doing this will reduce the number of false positives: fewer ebooks falsely identified as a subject category that does not apply. However, this will be at the cost of increasing the number of false negatives: fewer ebooks will be correctly identified as that subject category than otherwise would have been. If we increase the threshold for being classified as “Post-communism” for example, fewer documents will be *incorrectly* classified as having the subject “Post-communism” at the cost of missing some documents that *should* have been classified as “Post-communism”.

Choosing where to set the threshold is partly a matter of institutional policy. The LoC might decide to be quite permissive, for example, with subject tagging to drive discovery, or may decide to be more cautious in order to avoid irrelevant results.

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.

Details for this workflow are left intentionally high-level, without a great deal of specific information because the specifics of implementation depend on the outputs of two other experiments which we will carry out earlier in the Task Order. We will take the NLP experiment model which combines:

1. Best quality output for key metadata fields (Title, Author names, etc).
2. Ease of integration for adding additional visual or layout features to the model as part of an additional layer, additional set of features on the existing token/span objects, or as part of an additional pipeline step.

We will extend this NLP experiment model to use features such as:

- Text size
- Text position
- Position within the page
- Position within the overall document

To enhance and improve the outputs for fields where layout or style information are likely to be semantically relevant. For example, Title fields or Author names often have a distinction position and distinct formatting on the title page, which we can capture as features.

Note that implementation will vary depending on the library and approach chosen. Spacy, for example, among multiple other options, can be extended by adding custom attributes to tokens or spans which can be fed into the internal Tok2Vec trainable pipe, but can also be supplemented with custom pipeline steps which can wrap additional externally trained models from PyTorch, Tensorflow, Thinc, and so on, or via simple pipeline steps that filter the data before being passed to one of the existing pipeline components (Span Categorizer, Named Entity Recognition, etc). BERT based models can also be supplemented in multiple complementary ways.

The final decision about how to make use of layout information to supplement the NLP model will be made after we have trained and tested our NLP models.

We will:

1. Take the existing pre-trained NLP model (BERT or Spacy) from our earlier experiments
2. Supplement that model with layout information derived from the ePub or PDF files
3. Run the training cycle making use of this supplemental information
4. Run the optimization cycle / eval cycle (making use of the supplemental information)
5. Store the resulting model
6. Store Spacy (or BERT)'s metric data (output as part of the model evaluation)

After each of the models is complete and tested we will:

7. Run the  workflow across the *Test* set of ebooks (or the full set if we are using cross-validation) 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, Spacy DocBin data files (for corpora), Spacy models
- An instance of Digirati’s Django-based “Task Service” for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML.

# 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 as a whole are to help the Library answer the following research questions:

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

This particular data processing plan concerns, specifically, the question of:

> *…what technologies and workflow models are most promising to support metadata creation and assist with cataloging workflows?*

The goal is, for this model, to:

- Produce catalog metadata—primarily subject headings or genre terms—suitable for review by catalogers
- Provide, to the cataloger, information *about* the subject headings or genre terms that can assist them in choose the correct subject or genre term
- Provide, to the Library, for testing, a simple UI (such as a basic webform) to facilitate testing and review of the data by users

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:

- A lightweight UI for testing

- Structured data suitable for review by catalogers and other human users (rather than for automated metrics)
- 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)

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 and provide this bibliographic metadata in a form that can be reviewed by catalogers in a low-fidelity prototype in order to test the usefulness of machine generated subject or genre data in cataloging workflows.

Choosing the correct subject term was identified—alongside generating *new* subject terms for concepts not already covered by LCSH, which is outside the scope of this experiment—as a particular area of concern in the UX workshop from November 28th 2022.

In the case of this particular model, the expected scope is that the model will generate:

- Genre/form (using LCFGT as the vocabulary / thesaurus)
- Subject terms (using LCSH as the vocabulary / thesaurus)

For each ebook, this model  will provide a list of suggested subject or genre terms, along with the relevant identifier for that subject term, where available.

The model will also extract additional information which can assist the cataloger (the human-in-the-loop) in selecting one or more subject or genre terms as the likely best match.

This information could include:

- keywords or other terms that can assist in identifying the broad theme of this document
- quantitative information, such as the relative rank of this subject within the list of likely subjects identified by the ML workflow
- extractive or abstractive summaries of key sections of the document suitable for quick review by a cataloger
- information placing the subject or subjects identified by the ML workflow within a specific subject hierarchy, e.g. by displaying broader and narrower terms within the LCSH SKOS hierarchy so that the cataloger can select an appropriate term if the terms on screen are not the most promising or likely matches

The intent here is to supplement the most promising subject/genre producing workflow tested during the first 5 models with additional information in order to assist catalogers in selecting the correct identifier/term. Our expectation is that this would be one of:

- Model 2: Annif
- Model 3: Spacy
- Model 4: BERT

N.B. For this experiment, we may or may not train new models.

**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:

- Generated subject and genre terms for each ebook, with their accompanying identifiers in LCSH or LCGFT.
- Additional supplemental data such as:
    - keywords
    - abstractive or extractive summaries
    - subject or genre related information taken from taxonomies/source vocabulary data
- N.B. The interim data format will not be Marc but a simplified JSON representation that we can convert into Marc later.
- A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file.
    - We would expect this to be the same as the settings and parameters used in training the base model (Model 2, 3 or 4) used to provide the core subject and genre data for this experiment.
- Exports of the data models generated (where possible).
    - We would expect this to be the same as the model for the base model (Model 2, 3 or 4) used to provide the core subject and genre data for this experiment.
- 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)
- For this experiment we would not expect to produce any detailed metrics. The metrics would largely be a straight repetition of the metrics generated for the base model (2, 3 or 4) used to provide the subject data.
- Instead, we would expect to provide record-level data for every eBook and a simple user interface to allow cataloguers to review this record-level data alongside the generated data for user testing and review purposes.

**A4: Description of intended use**

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

The experiment will reuse trained models on the ebook plaintext and additional models (such as abstractive summarisation or keywording tools) will be used to generate additional supplemental data.

The primary intended use for the data generated is as part of the final report, and for testing by catalogers and other end users,  rather than for further use in a production context.

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset's technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains.<br>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.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs<br>c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records<br>b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance | The information in this dataset originated from four collections of LoC ebooks: |

| | |
|---|---|
| 1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | 1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br><br>Further details to be provided by LoC. |
| d) Compilation methods<br>1. How is/was this dataset compiled, when, and by whom?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br><br>Further details to be provided by LoC. |
| e) Preprocessing steps<br>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>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:<br>a. Conversion of PDF and epub to plaintext using a mixture of tools, including Grobid, PDFAlto, and Calibre<br>b. Normalization of character set encoding (conversion to unicode for files that aren’t encoded as unicode, if any)<br>c. Normalize whitespace<br><br>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 pipeline stages in the subject or genre generation pipelines.<br><br>For example, we actively do not want to remove stop words or lemmatize the text if our goal is to provide a human readable summary of the introduction to the ebook, for example. We *will* want to remove stop words and lemmatize the |

<table>
<tr><td></td><td>text, if we are generating subject or genre information.<br><br>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)..<br><br>3. The metadata is organized as MarcXML files following usual LoC cataloging practice.</td></tr>
<tr><td colspan="2">f) Potential risks to people, communities and organizations &amp; strategies for risk mitigation:<br>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.)<br>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.)</td></tr>
<tr><td colspan="2">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.<br><br>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.</td></tr>
<tr><td colspan="2">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?</td></tr>
<tr><td colspan="2">Not in scope. 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.<br><br>The records will not, as part of this experiment, be made public.</td></tr>
<tr><td>h) Copyright, licensing, rights, and/or privacy restrictions<br>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.</td><td>The material comprises a mixture of open access and copyrighted ebooks.<br><br>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.</td></tr>
</table>

*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.

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

We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. We may also split out digitized from born-digital material at the test stage, in order to have comparative data.

However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.

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

This experiment uses Spacy, a widely used NLP library, or may use BERT, depending on the most promising subject/genre tagging workflow tested as part of the earlier 5 models. As such, this section will repeat the same information on the other data processing plan(s).

See:

- Model 2: Annif
- Model 3: Spacy
- Model 4: BERT

Spacy and/or BERT pretaining data include the following:

The core English models are pre-trained on OntoNotes 5, ClearNLP Constituent-to-Dependency Conversion (Emory University), WordNet 3.0 (Princeton University) and Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl) (Explosion) datasets. We would expect to use the large English language model, and the transformer based model for English records.

| |
|---|
| RoBERTa, is additionally trained on the RoBERTa base dataset (see also https://huggingface.co/roberta-base). |
| 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.<br>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. |
| N/A |

## 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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | Assisted Subject / Genre Cataloging |
| b) Intended use | Automated extraction and supplementation of subject/genre classifications from ebooks for use by catalogers in human-in-the-loop workflows. |
| c) Limitations | The primary input is plaintext only.<br>Visual information (size, font style, location on page, location within the book structure, etc) present in the ebooks is out of scope for this experiment (although will be tested in a different experiment).<br>Both Spacy and BERT  benefit from smaller blocks of text, so we may need to select a subset of the ebooks (the first N thousand words, for example), or process the books in chunks. |

<table>
<tr><td>d) Copyright and licensing details for the model</td><td>Spacy is licensed under the MIT license, see: https://github.com/explosion/spaCy/blob/master/LICENSE Similarly BERT and derivative models are generally released under open source licenses. https://github.com/google-research/bert/blob/master/LICENSE</td></tr>
<tr><td>e) Link to documentation</td><td>https://spacy.io/ and/or https://github.com/google-research/bert and https://huggingface.co/docs/transformers/model_doc/distilbert and https://huggingface.co/docs/transformers/model_doc/roberta</td></tr>
<tr><td>f) Predicted performance metrics (range)</td><td>N/A<br><br>This experiment is not designed to produce performance metrics at scale. Instead the experiment is designed to provide data to users (human users) for review.</td></tr>
<tr><td>g) Actual performance metrics</td><td>N/A</td></tr>
<tr><td>h) Audit schedule (how often and how many times will performance metrics be checked?)</td><td>N/A</td></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td colspan="2">N/A<br><br>Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report. The primary goal is to provide a user interface and document-level data that can be tested and reviewed with LoC users rather than to provide evaluable metrics at scale.</td></tr>
<tr><td colspan="2">i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning.</td></tr>
<tr><td colspan="2">There are three possible base models we will use to extract the initial subject data:<br><br>● Model 2: Annif<br>● Model 3: Spacy<br>● Model 4: BERT<br><br>In addition to the workflows outlined in these documents we will:<br><br>1. Take plaintext extracted from PDFs and ePubs stored in an Amazon S3 bucket<br>2. Run the most promising subject/genre cataloging model (evaluated as part of the earlier testing of models)<br>3. Supplement the data from these models with:<br>a. Keywords: we would expect to test multiple approaches to this as part of the data generation, including:</td></tr>
</table>

        i. Using Spacy’s approach to identifying keywords directly
        ii. TF-IDF (term frequency inverse document frequency)
        iii. Topic-term matrices generated by BERTopic using cTF-IDF: this approach clusters the documents before generating TF-IDF data. See: https://maartengr.github.io/BERTopic/algorithm/algorithm.html#5-topic-representation

    N.B. These keywords and other clustering data are generated already as part of the ML workflow that looks for subject headings. However, in this case, we would explicitly generate or surface this information for review by catalogers

    b. Summarization: again, we would expect to test several approaches to generating summary data (taken from the text of the document) which can be shown to the cataloger, including:
        i. PyTextRank or other Spacy based approaches
        ii. Gensim summarization
        iii. Transformer based models abstractive or extractive models for summarization such as:
            1. BART https://huggingface.co/facebook/bart-large-cnn
            2. Distillbart, etc
    c. Data taken from the LCSH taxonomy for each of the candidate subjects such as broader or narrow terms to help the user select the appropriate term

After this data is complete we will:

4. Store any supplemental information that we pre-generate (N.B. some of the data might be fetched at access time, e.g. by a call to an LoC API)
5. Create a lightweight UI to allow users to:
    a. review the candidate subjects alongside the supplemental information
    b. review the original MarcXML record to compare to the generated data

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, Spacy DocBin data files (for corpora), Spacy models
- An instance of Digirati’s Django-based “Task Service” for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML.
- A lightweight Django-Rest-Framework based API and accompanying template based frontend that provides access to the data for human-in-the-loop review

N.B. our expectation is that this particular UI will be low-fidelity: a simple web form or similar.

# 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 as a whole are to help the Library answer the following research questions:

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

This particular data processing plan concerns, specifically, the question of:

> *…what technologies and workflow models are most promising to support metadata creation and assist with cataloging workflows?*

The goal is, for this model, to:

- Produce catalog metadata—primarily information about people (authors, editors, etc.) and/or similar *entities* such as places or organizations such as publishers—suitable for review by catalogers
- Provide, to the cataloger, supplemental information such as:
    - a list of possible LCNAF or other authority matches for a given named person (author, editor, etc.)
    - data extracted from the text or machine generated record for the object such as keywords, related places or person names, dates, etc
    - surrounding textual context
- With the aim being to provide the user/cataloger with information to assist in disambiguating entities (people, places, organizations) and *linking* them to their related authority controlled identifier(s)
- Provide, to the Library, for testing, a simple UI (such as a basic webform) to facilitate testing and review of the data by users

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:

- A lightweight UI for testing
- Structured data suitable for review by catalogers and other human users (rather than for automated metrics)
- 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)

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 and provide this bibliographic metadata in a form that can be reviewed by catalogers in a low-fidelity prototype in order to test the usefulness of machine generated data in assisting catalogers to match entities within the document data to their correct identifiers.

Disambiguating between multiple potential matching identifiers in controlled —authors with multiple matching LCNAF entries for people with the same name, for example—was identified as a particular area of concern in the UX workshop from November 28th 2022. Catalogers described the process of seeking additional information about the potential authors and/or additional information from within the document to identify which of the potential named individuals was the correct one as a time-consuming process. The aim of this experiment is to partly automate some of this process to evaluate whether this assists the cataloger.

In the case of this particular model, the expected scope is that the model will generate:

- Author, editor, or other person data via token classification from within the document text
- Organization data (such as Publisher or rights holder information) from within the document text
- Place information (such a Place of Publication) from within the document text

The focus of the experiment is likely to be more on Person data than Place or Organization data, but it would be interesting to compare approaches across the three.

For each ebook, this model  will provide a list of suggested named people (or places, or organizations) along with the appropriate identifier, for example, LCNAF id, for each.

The model will also extract additional information which can assist the cataloger (the human-in-the-loop) in selecting one or named entities as the likely best match.

This information could include:

- keywords or other terms that can assist in identifying the broad theme of this document
    - For example, it might be useful to know that the document concerns "physics", especially if this can be matched to, for example, LCNAF information such as Field of Activity or Occupation
- extracts from the text surrounding the entity (Person, Place, Organization) so that the entity can be identified in context
- lists of other entities (Places, People, Organizations) from within the document, for example, if there are 3 possible matches, and one is affiliated with the University of Nairobi, it would potentially be useful to know that Kenya and Nairobi are both commonly named places within the document text, or commonly co-occur *within the text* with this person's name.
- data pulled from the appropriate authority records (LCNAF, etc.) to show to the cataloguer in order to assist in disambiguation. This could involve comparing key terms in the authority record to key terms from within the document text, for example.

The intent here is to supplement the most promising token classification workflow tested during the first 5 models with additional information in order to assist catalogers in selecting the correct identifier/term. Our expectation is that this would be one of:

- Model 3: Spacy
- Model 4: BERT

In particular, the token classification (entity recognition) elements of each of these workflows, rather than the text classification (subject, genre) elements of these workflows.

N.B. For this experiment, we may or may not train new models, but we would expect to potentially simplify some of the pipelines to focus on just those entities/terms/tokens that are relevant to this experiment.

**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:

- Generated authority controlled identifiers for People, Places, Organizations, etc.
- Additional supplemental data such as:
    - keywords
    - other related entities
    - data taken from authority files (LCNAF extracts, etc)
- N.B. The interim data format will not be Marc but a simplified JSON representation that we can convert into Marc later.
- A record of any configuration, hyperparameters or other settings used in training the model as a machine readable file.
    - We would expect this to be the same as the settings and parameters used in training the base model (Model 3 or 4) used to provide the core author, editor, publisher, etc. information.

| |
|---|
| ● Exports of the data models generated (where possible).<br>○ We would expect this to be the same as the model for the base model (Model, 3 or 4) used to provide the core data for this experiment.<br>● 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)<br>● For this experiment we would not expect to produce any detailed metrics. The metrics would largely be a straight repetition of the metrics generated for the base model (2, 3 or 4) used to provide the subject data.<br>● Instead, we would expect to provide record-level data for every eBook and a simple user interface to allow cataloguers to review this record-level data alongside the generated data for user testing and review purposes. |
| **A4: Description of intended use** |
| *Please describe how the data will be used in the experiment.*<br>The experiment will reuse trained models on the ebook plaintext and additional models (such as abstractive summarisation or keywording tools) will be used to generate additional supplemental data.<br>The primary intended use for the data generated is as part of the final report, and for testing by catalogers and other end users,  rather than for further use in a production context. |

## 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 | *Task Order 1 ebook dataset* |
| b) Composition<br>1. Please describe the dataset’s technical composition, including file type, content type, number of items, and relative size.<br>2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains.<br>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 | The dataset consists of ebooks and MarcXML files with catalog records for those ebooks.<br>1. Technical composition:<br>a. Total number of items: 23130<br>b. File type:<br>i. 13070 PDFs<br>ii. 10060 epubs<br>c. Content type: ebooks<br>d. Relative size: ~250GB<br>2. Full data audit to follow.<br>a. Languages (28 languages):<br>i. English ~18,000 records<br>ii. German ~700 records<br>iii. Other: ~ 4,000 records |

| | |
|---|---|
| or it may be a series of folders containing images derived from a particular source. | b. Genre: Approx 11% of the records have a listed genre. For details see full data audit.<br>c. Period: 21st century ebooks. For details see full data audit.<br>3. The dataset comprises four discrete sub-collections:<br>a. CIP (13802 items)<br>b. Open access (5835 items)<br>c. E Deposit ebooks (403 items)<br>d. Legal reports (3750 items)<br>Each collection is organized as a folder of ebooks in PDF or ePub format.<br>Accompanying each folder is a single MarcXML file containing the catalog records for each of the ebooks within that sub-collection. |
| c) Provenance<br>1. Where did the information in this dataset originate? Please include relevant links where possible.<br>2. Include any version information if available. | The information in this dataset originated from four collections of LoC ebooks:<br>1. Ebooks provided as part of CIP prepublication cataloging<br>2. Ebooks provided as part of E-Deposit registration<br>3. Ebooks provided as part of the Open Access ebooks program<br>4. Legal reports<br>Further details to be provided by LoC. |
| d) Compilation methods<br>1. How is/was this dataset compiled, when, and by whom?<br>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).<br>2. The files were uploaded to a private Amazon S3 bucket provisioned by Digirati for data storage for the Task Order 1 experiment.<br>Further details to be provided by LoC. |
| e) Preprocessing steps | 1. The dataset comprises a mixture of PDFs and epub files. The preprocessing steps for this experiment were: |

<table>
<tr>
<td>1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment?<br>2. How was material selected for inclusion or exclusion in the dataset?<br>3. Is the data organized according to a schema, content standard, or other standard? If yes, which one?</td>
<td>a. Conversion of PDF and epub to plaintext using a mixture of tools, including Grobid, PDFAlto, and Calibre<br>b. Normalization of character set encoding (conversion to unicode for files that aren’t encoded as unicode, if any)<br>c. Normalize whitespace<br><br>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 pipeline stages in the subject or genre generation pipelines.<br><br>For example, we actively do not want to remove stop words or lemmatize the text if our goal is to provide a human readable summary of the introduction to the ebook, for example. We <i>will</i> want to remove stop words and lemmatize the text, if we are generating subject or genre information.<br><br>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)..<br><br>3. The metadata is organized as MarcXML files following usual LoC cataloging practice.</td>
</tr>
<tr>
<td colspan="2">f) Potential risks to people, communities and organizations &amp; strategies for risk mitigation:<br>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.)<br>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.)</td>
</tr>
</table>

| |
|---|
| 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.<br><br>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? |
| Not in scope. 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.<br><br>The records will not, as part of this experiment, be made public. |

| | |
|---|---|
| h) Copyright, licensing, rights, and/or privacy restrictions<br>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.<br><br>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. |

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

We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. We may also split out digitized from born-digital material at the test stage, in order to have comparative data.

However, we would expect that for the majority of the experiment the dataset will be split randomly and any specific ebook (and associated MarcXML) could be used for training, validation or test.

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

This experiment uses Spacy, a widely used NLP library, or may use BERT, depending on the most promising token classification workflow tested as part of the earlier 5 models. As such, this section will repeat the same information on the other data processing plan(s).

See:

- Model 3: Spacy
- Model 4: BERT

Spacy and/or BERT pretaining data include the following:

The core English models are pre-trained on OntoNotes 5, ClearNLP Constituent-to-Dependency Conversion (Emory University), WordNet 3.0 (Princeton University) and Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl) (Explosion) datasets. We would expect to use the large English language model, and the transformer based model for English records.

RoBERTa, is additionally trained on the RoBERTa base dataset (see also https://huggingface.co/roberta-base).

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. |
|---|
| N/A |

## 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.

| **D1: Machine Learning or Artificial Intelligence Model** | |
|---|---|
| a) Model Details | Assisted Token Classification and Entity Disambiguation |
| b) Intended use | Automated extraction of people, places and organizations from ebooks for use by catalogers in human-in-the-loop workflows. |
| c) Limitations | The primary input is plaintext only.<br>Visual information (size, font style, location on page, location within the book structure, etc) present in the ebooks is out of scope for this experiment (although will be tested in a different experiment).<br>Both Spacy and BERT  benefit from smaller blocks of text, so we may need to select a subset of the ebooks (the first N thousand words, for example), or process the books in chunks. |
| d) Copyright and licensing details for the model | Spacy is licensed under the MIT license, see: https://github.com/explosion/spaCy/blob/master/LICENSE Similarly BERT and derivative models are generally released under open source licenses. https://github.com/google-research/bert/blob/master/LICENSE |
| e) Link to documentation | https://spacy.io/ and/or https://github.com/google-research/bert and https://huggingface.co/docs/transformers/model_doc/distilbert and https://huggingface.co/docs/transformers/model_doc/roberta |
| f) Predicted performance metrics (range) | N/A |

| | |
|---|---|
| | This experiment is not designed to produce performance metrics at scale. Instead the experiment is designed to provide data to users (human users) for review. |
| g) Actual performance metrics | N/A |
| h) Audit schedule (how often and how many times will performance metrics be checked?) | N/A |

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.

N/A

Our aim will be to gather as much information as possible to feed into discussions with LoC and the final report. The primary goal is to provide a user interface and document-level data that can be tested and reviewed with LoC users rather than to provide evaluable metrics at scale.

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

There are two possible base models we will use to extract the initial token/entity data (primarily people, but also places and organizations):

- Model 3: Spacy
- Model 4: BERT

In addition to the workflows outlined in these documents we will:

1. Take plaintext extracted from PDFs and ePubs stored in an Amazon S3 bucket
2. Run the most promising token classification model (evaluated as part of the earlier testing of models)
3. Supplement the data from these models with:
    a. Keywords: we would expect to test multiple approaches to this as part of the data generation, including:
        i. Using Spacy's approach to identifying keywords directly
        ii. TF-IDF (term frequency inverse document frequency)
        iii. Topic-term matrices generated by BERTopic using cTF-IDF: this approach clusters the documents before generating TF-IDF data. See: https://maartengr.github.io/BERTopic/algorithm/algorithm.html#5-topic-representation

    N.B. These keywords and other clustering data are generated already as part of the ML workflow that looks for subject headings. However, in this case, we would explicitly generate or surface this information for review by catalogers in the context of identifying authors, editors, publishers, etc. *This would also differ from Assisted Cataloging 1, as we would potentially generate similar keyword information from authority file records, too.*

    b. Surrounding textual context
    c. Data taken from the LCNAF or other authority data files

   d. Lists of additional co-occurring entities within the document such as Places or Organizations that typically accompany this Person within the document. If a potential matching author has an affiliation with a particular Place or Organization, or commonly co-occurs with another named Person, this would be useful information to help the cataloger choose the correct LCNAF entry.

After this data is complete we will:

4. Store any supplemental information that we pre-generate (N.B. some of the data might be fetched at access time, e.g. by a call to an LoC API)
5. Create a lightweight UI to allow users to:
   a. review the candidate authors, editors, publishers, places of publication, etc. alongside the supplemental information
   b. review the original MarcXML record to compare to the generated data

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, Spacy DocBin data files (for corpora), Spacy models
- An instance of Digirati's Django-based "Task Service" for queuing up long-running jobs such as text preprocessing, or data reformatting, running on the same AWS estate as the EC2 instances for ML.
- A lightweight Django-Rest-Framework based API and accompanying template based frontend that provides access to the data for human-in-the-loop review

N.B. our expectation is that this particular UI will be low-fidelity: a simple web form or similar.