LIBRARY LIBRARY OF CONGRESS

United Nations Learning Session
September 12, 2024

Responsible and Ethical Use of AI in Libraries and Archives

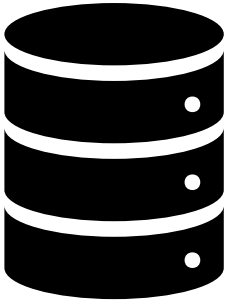
Exploring Computational Description: Experiment Results

Caroline Saccucci and Abigail Potter

Elements of Machine Learning

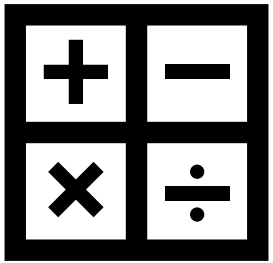
Processes that are trained recognize and predict patterns in data

Data



- Our/Your content
- Data readiness
- Training data
- Tuning data
- Validation data
- Target data
- Output data

Model

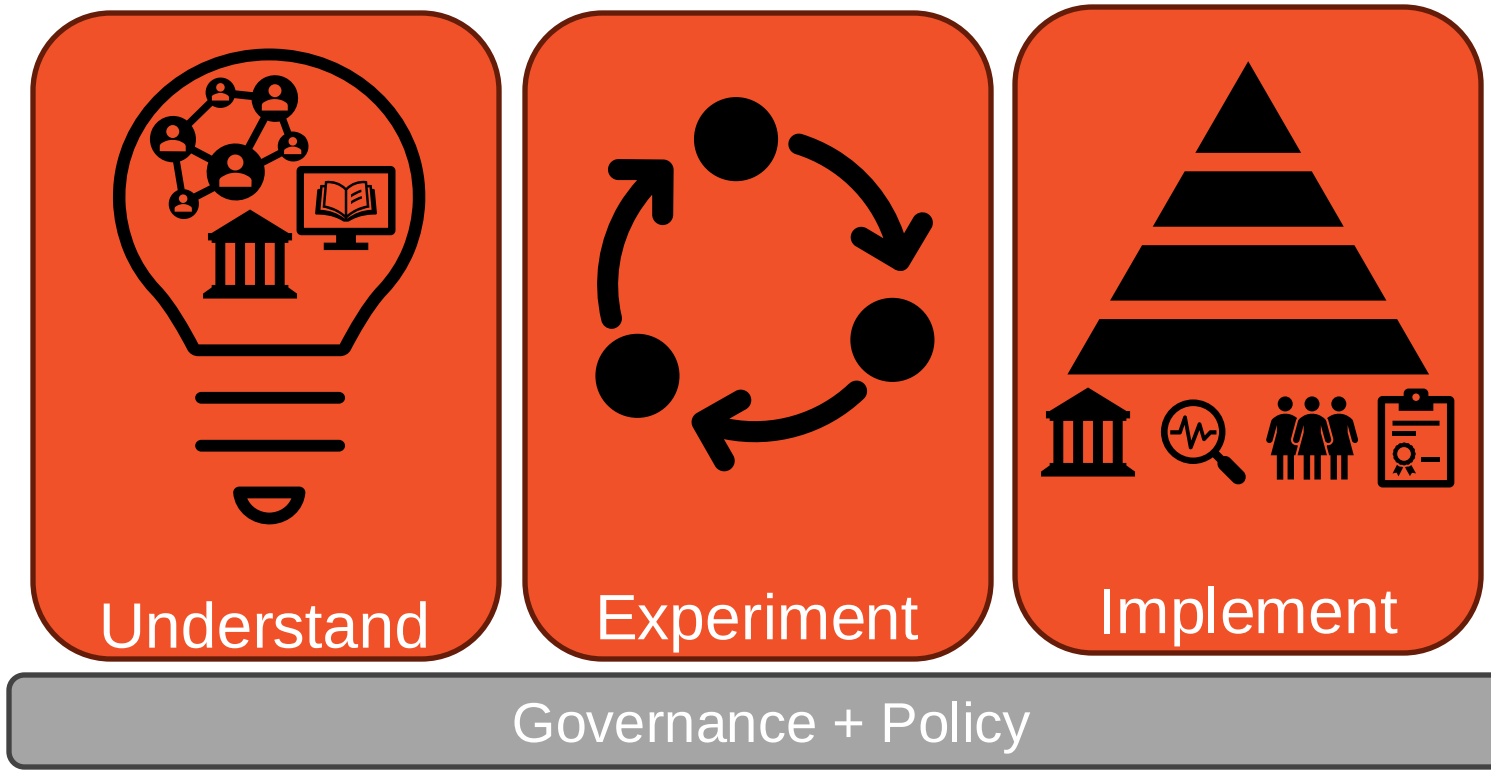


- End to end workflow, pipeline, platform or tool
- Architectures
- Type of training
- Libraries utilized
- Frameworks or platforms

People

- Develop use cases
- Represented in the data
- Design & sell AI systems
- Impacted by AI systems
- Evaluate & implement AI systems
- Responsible for AI outputs

LC Labs AI Planning Framework : Phases

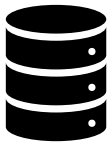


LC Labs AI Planning Framework : Phases & Activities

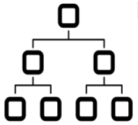
Understand

Risks & benefits

Principles & values

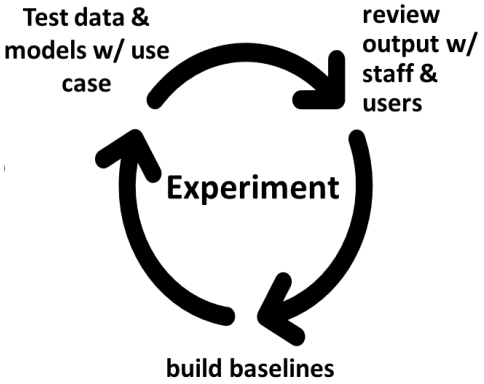


Data readiness



Needs & expertise

Experiment

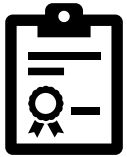


Implement

Strategy & Roadmap

Skills & Capacities

Monitoring & Measuring



Shared Quality Standards

Evaluations & Documentation

Collaborate & Assess

Resources & Standards

Governance + Policy

Example Data Processing Plan

Attachment J2 - Data Processing Plan Template

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

Section A: General (required)

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

Section B: Data Documentation (required)

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

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

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

1) Please describe the purpose of this dataset with relation to the ML/AI workflow. Explicitly address if it is being used as training, validation, or test data.
The dataset will be explicitly split into training, validation and test data without cross-evaluation. The split will be random, and follow a standard 80/10/10 split. We would expect the training, validation, and test_data to comprise randomly assigned examples from all four of the sub-divisions (CIP, OA, E Deposit, Legal Reports) within the dataset. We may split the dataset by language to evaluate specific language models, for example, using a German language base language for German texts. However, the overall volume of non-English material is low, so this may not be required. A small subset of the training data split will be used for few-shot learning, or prompt tuning.
b) For training data: 1) if the model is pre-trained, describe the data on which it was trained; 2) if the model will be fine-tuned, outline the data involved in this process; 3) if the model is being trained from scratch, outline the plan for creating training data.
Each of the large language models that might be evaluated has been trained on its own dataset, and in some cases, the precise details of the training dataset is left unclear or deliberately held back for competitive advantage. In some cases, models may potentially be trained on copyright or non-public-domain information. However, the broad datasets tend to be the same for most models. For example, LLama-2 is trained on (information from wikidata): • Webpages scraped by CommonCrawl • Open source repositories of source code from GitHub • Wikipedia in 20 different languages • Public domain books from Project Gutenberg • The LaTeX source code for scientific papers uploaded to ArXiv • Questions and answers from Stack Exchange websites
and additionally fine-tuned using 27,540 prompt-response pairs created for Llama-2 and reinforcement learning with human feedback (RLHF) was used with a combination of 1,418,091 Meta examples and seven smaller datasets. Similarly, Google say, for Gemma, that: These models were trained on a dataset of text data that includes a wide variety of sources, totaling 6 trillion tokens. Here are the key components: Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content. Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions.

Research questions & goals

ECD1

- Test multiple methods with ebook data
- Understand performance baselines
- Initial understanding of data quality

ECD2

- Provide more ebook data for training and tuning models
- Output in valid MARC
- Prototype HITL catalog assistance workflows

Data

ECD1

- Ground Truth data for testing and validation
  - CIP (13802)
  - Open Access (5835)
  - E Deposit (403)
  - Legal Reports (3750)
  - Plus, associated catalog records
- **Key metadata**: author, title, creation date, issuance date, form/genre, subject, LCCN, and ISBN
- Predominately English language, some German and Spanish, ebooks in epub and pdf formats
- Did not select the training data set to be balanced or representative across subjects or genres

ECD2

- 119,823 unique CIP ebooks; ~1TB
- Output in valid MARC – Structured data

010: Library of Congress Control Number (LCCN)
020: International Standard Book Number (ISBN)
050: Call Number
082: Dewey Decimal Classification Number
100: Main Entry - Personal Name
245: Title Statement
264: Production, Publication, Distribution, Manufacture, and Copyright Notice
600: Subject: Personal Name
650: Subject Added Entry - Topical Term
651: Subject: Geographic Name
655: Genre
700: Added Entry - Personal Name

What was tested

ECD1

- **Models:** Bert, Spacy, GPTs with variations (NLP, NER, LLMs, transformer and non-transformer)
  - Token classification
  - Text classification
  - Data serialization
- **Human in the Loop (HITL) workflows**: combining AI output and human review or verification.

ECD2

- **HITL prototypes** for reviewing output
- **Open source LLMs** – primarily MistralAI - 30 experiment runs
  - **LLM Prompting**
  - **LLM Fine-tuning**
- Vector “search” to match field values to authority records

Results

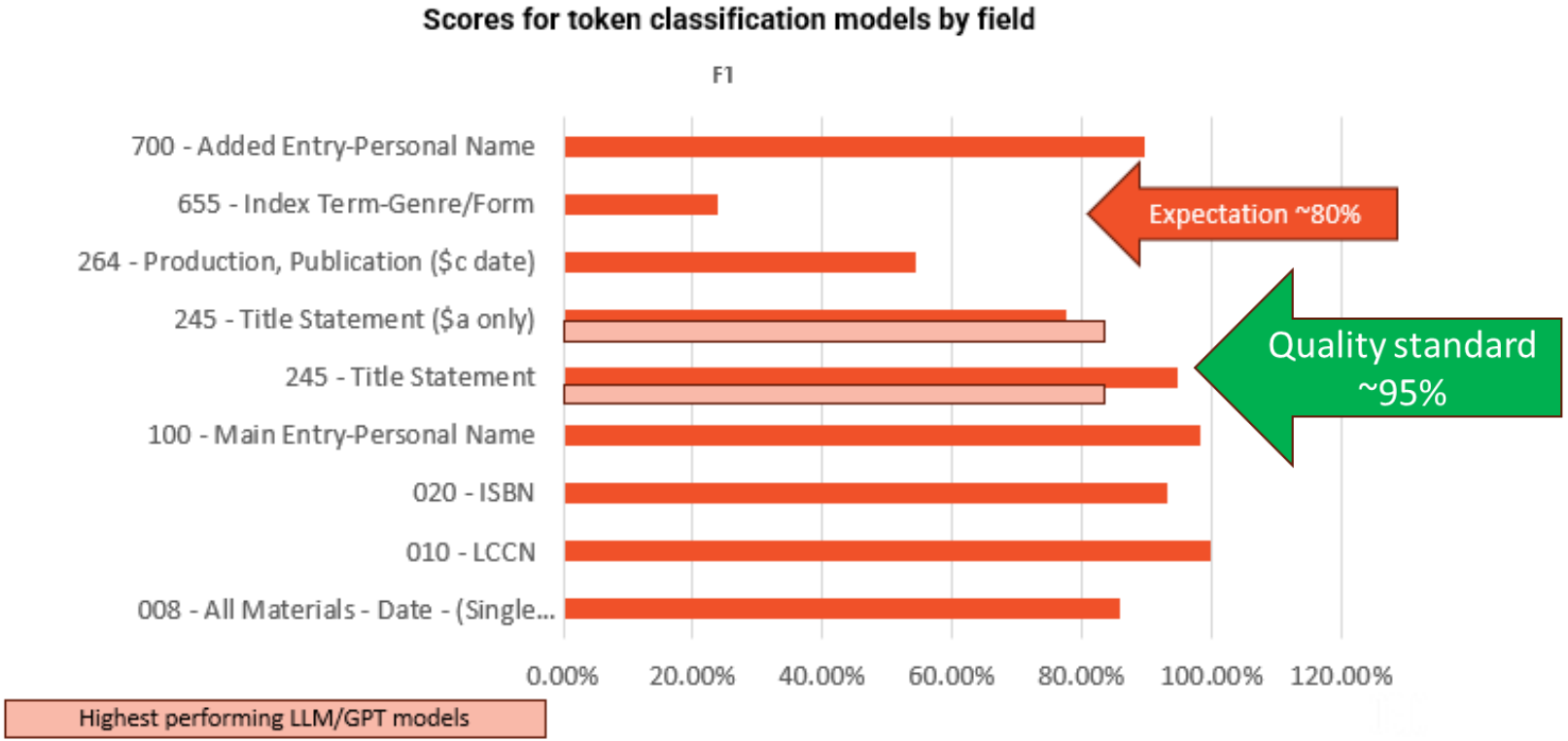
ECD1

- **Subject Classification is challenging**
  - The number of **instances** of each individual subject are very low, with most subjects only appearing once in the entire corpus
  - A very small number of **subjects** appear many times
  - Subjects are very unbalanced across the entire dataset
- Evaluation of the user facing assisted cataloging prototypes suggested that:
  - **Catalogers are receptive to automated suggestions**
  - Use of authority data was valuable
  - Review of data *in-context* was valuable
  - More work is needed to produce *full* bibliographic via automated methods

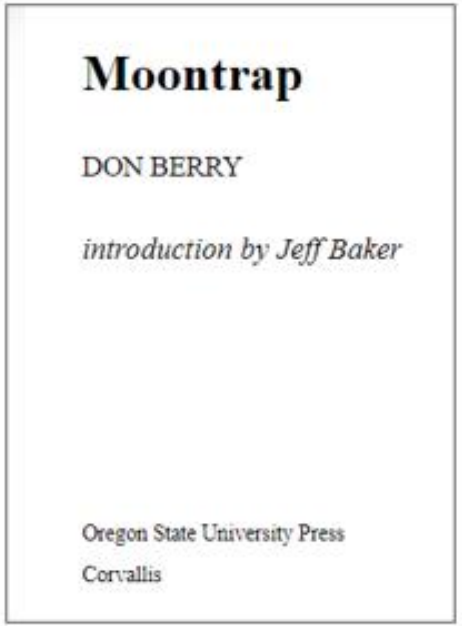
ECD2

- Producing valid MARC records using machine methods is possible
- Overall **accuracy ~80+%** for most fields and subfields
- (6xx) **Subject fields were accurate 46%** of the time
- LLMs can be constrained to produce:
  - Structure data
  - Subfield level data
- Fine-tuned LLMs generally perform better than other options

ECD1: Results: Text Classification, sample



ECD1: Results: Text Classification, sample

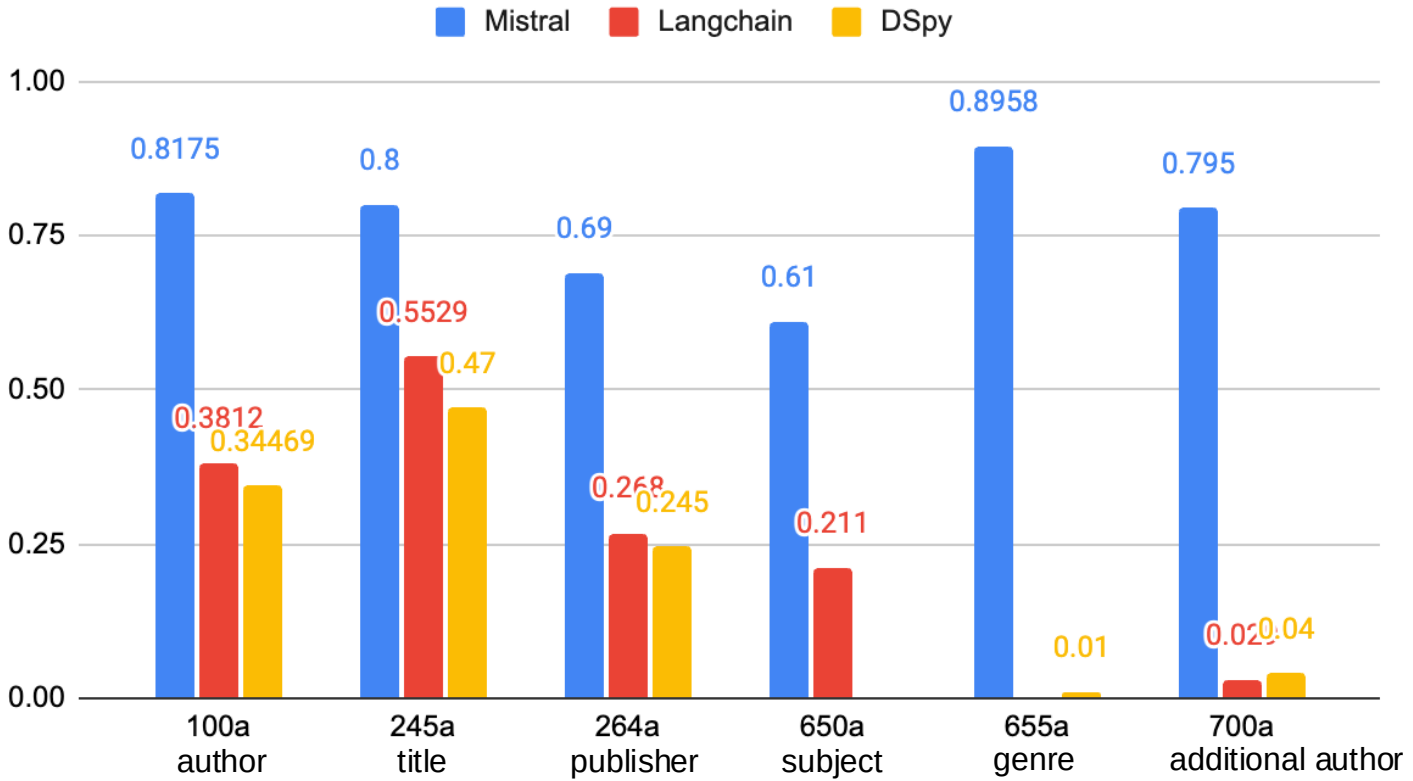


Results after applying Annif

Green shaded areas are exact match to MARC XML

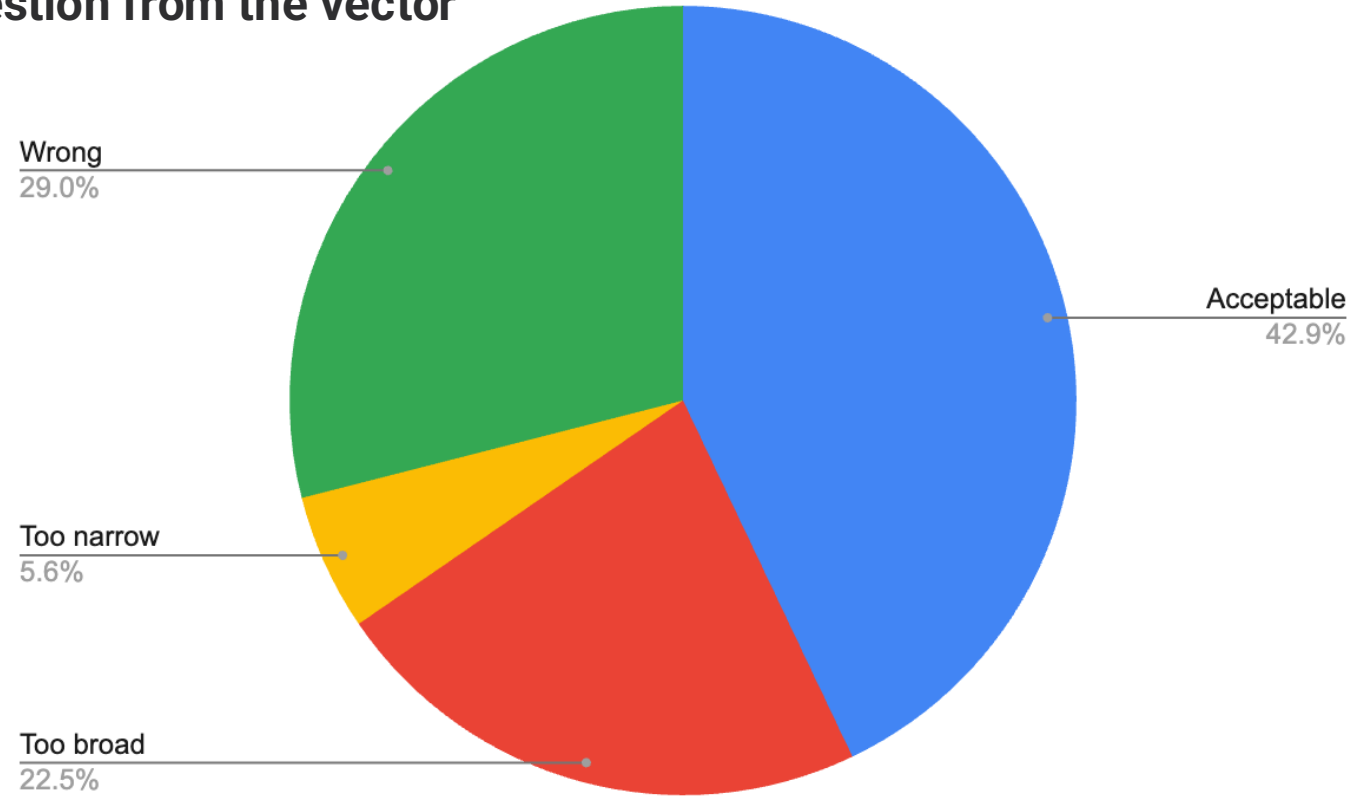
<http://id.loc.gov/authorities/subjects/sh85013380>	Berries
<http://id.loc.gov/authorities/subjects/sh2008107177>	Married people--Fiction
<http://id.loc.gov/authorities/subjects/sh97002963>	Cyberspace
<http://id.loc.gov/authorities/subjects/sh2009125638>	Fur trade--Fiction
<http://id.loc.gov/authorities/subjects/sh2008103545>	Farm life--Fiction
<http://id.loc.gov/authorities/subjects/sh88007452>	Beats (Persons)
<http://id.loc.gov/authorities/subjects/sh93007756>	Shoshoni women
<http://id.loc.gov/authorities/subjects/sh2002004972-781>	Oregon--Clackamas River Valley
<http://id.loc.gov/authorities/subjects/sh2008112706>	Trappers--Fiction
<http://id.loc.gov/authorities/subjects/sh85146781>	Willamette River (Or.)

ECD2: Results for core MARC fields



ECD2: Results of manual subject review

Top ranked suggestion from the vector search



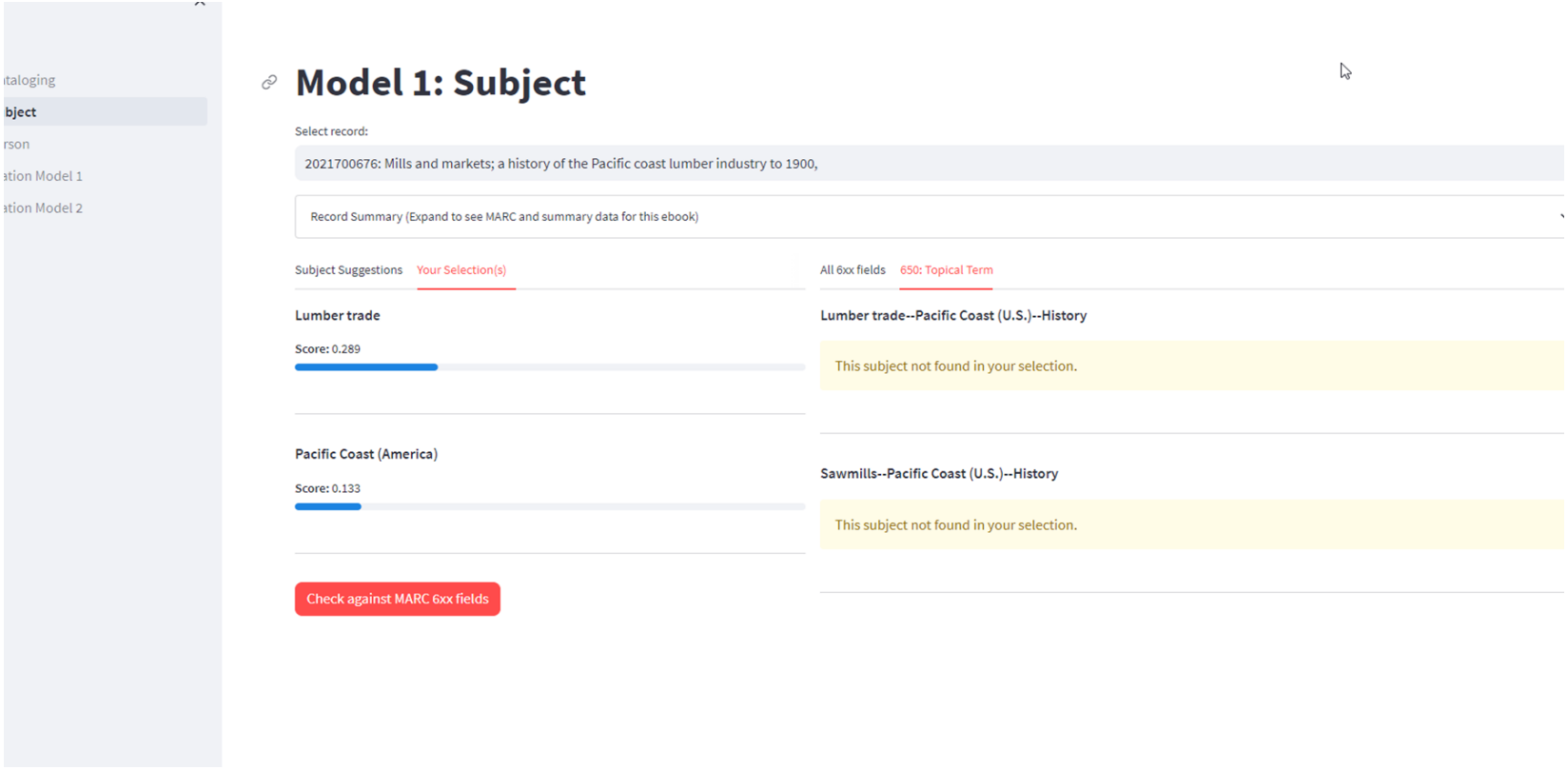
ECD2: Subject review comments

~250 comments on subject predictions assessed to be “wrong” by reviewers.

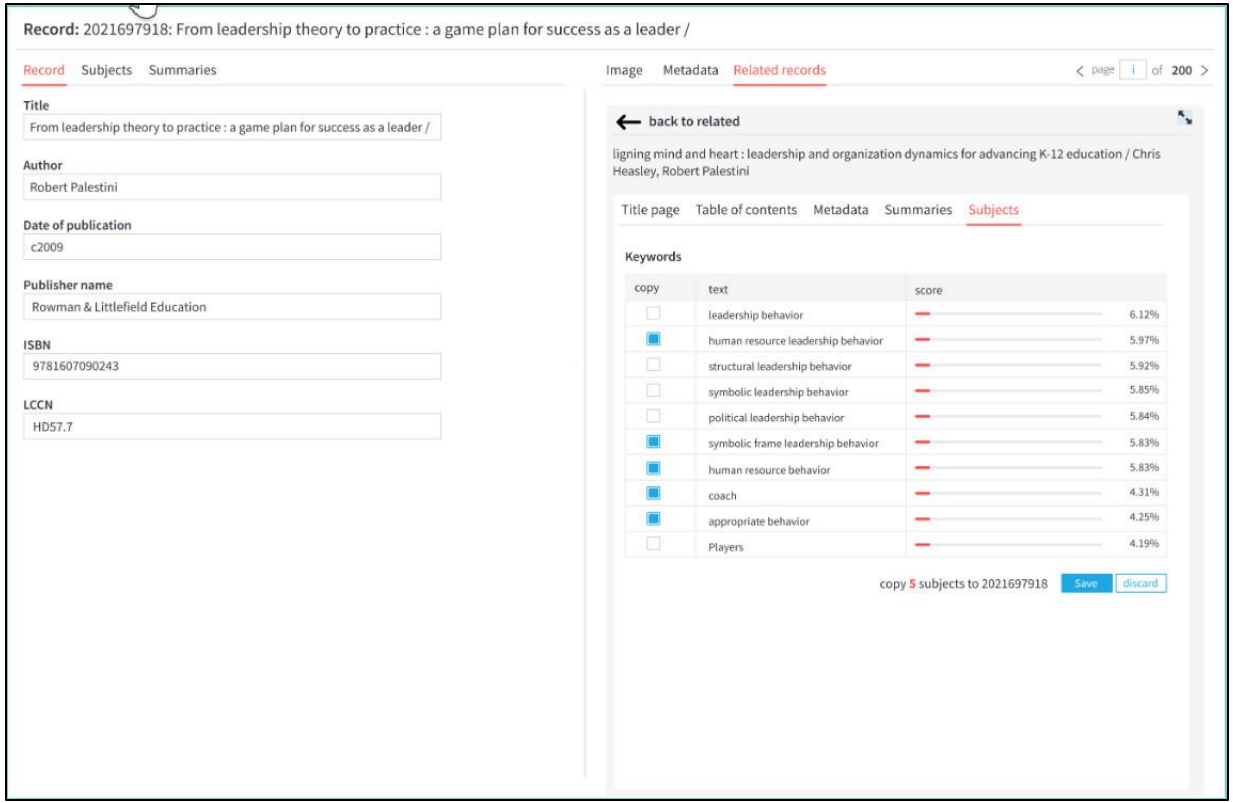
Typical patterns of comments, however:

- Wrong subdivision order
- Subjects being too broad, as, for example, there needed to be a geographic subdivision
- Subjects being too narrow, as, for example, when the geographic subdivision didn’t include all of the places covered by the work
- Incorrect MARC field, e.g. when a term that should be 610 was predicted for 650, etc.
- Subdivisions being provided alone rather than the entire subject

ECD1: Assisted Cataloging HITL Prototype



ECD2: Assisted Cataloging HITL Workflows



ECD2: Assisted Cataloging HITL Workflows

Subject Added Entry - Topical Term

Television broadcasting -- History -- 20th century. -- Europe, Eastern

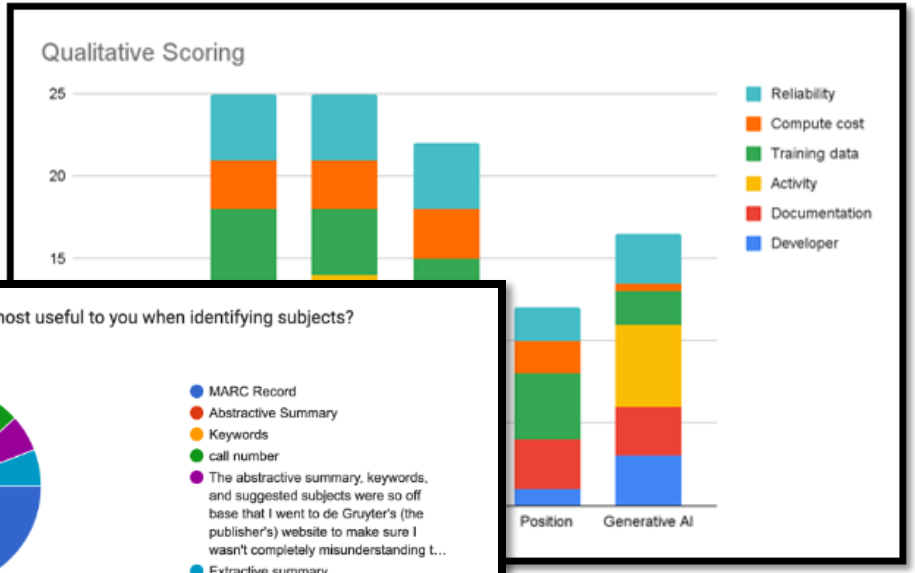
MARC=650 \\$aTelevision broadcasting$xHistory$y20th century.$zEurope, Eastern

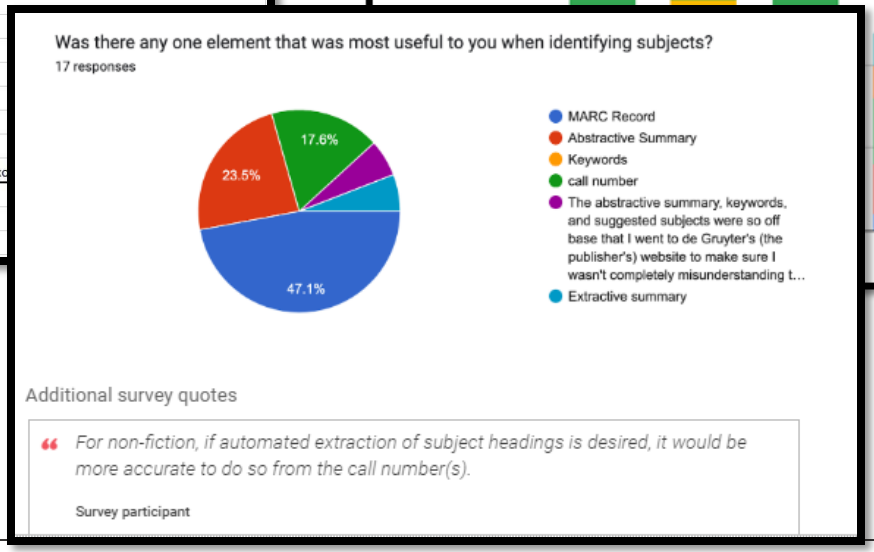
LCSH	Text	Acceptable	Too Broad	Too Narrow	Wrong	
sh2010116008	Television broadcasting--History	☐	☐	☐	☐	💬
sh2010116029	Television broadcasting--Soviet Union	☐	☐	☐	☐	💬
sh2008112751	Television broadcasting--Europe	☐	☐	☐	☐	💬
sh85133505	Television broadcasting--Bibliography	☐	☐	☐	☐	💬
Original Prediction	Television broadcasting -- History -- 20th century. -- Europe, Eastern	☐	☐	☐	☐	💬

Previous

Assessing AI Outcomes

Cataloging Field	Framework/Model		
Token Classification		F1 score - average of precision and recall scores	
LCCN/010	Hugging Face	100%	
Personal Name/100	Spacy	99%	*performance standard for this task?*
Title/245	Spacy	98%	
Added Name/700	Spacy	94%	
ISBN/020	Hugging Face	83%	
All fields	Spacy: RoBerta	80%	expect ~80% accuracy
Title & Author	GPT 3.5	76%	
Title & Author	Llama-2	76%	
Production/264	Spacy	75%	
All fields	HF: Distilbert-Base	74%	
All fields	Spacy: LEV	74%	
Series Statement/490	Hugging Face	71%	
Title & Author	Hugging Face	63%	
Title & Author	Spacy+HF	57%	
Title & Author	Spacy	56%	
Text Classification			exc
Subject Classification	Annif - MLLM	18%	
Subject Classification	Annif - NNE	15%	
Subject Classification	Annif - Ensemble	13%	





Assessing AI Outcomes

Responsible	Effective	Practical
Balance risks & benefits Compliant Supports LOC AI principles*	Tested w/ LOC data Reviewed by LOC Meets standards	Cost effective Integrate w/ LOC Stable over time

Caroline's hot takes

1. More and carefully constructed training data is needed
   ½ of the training data contained similar patterns of LCSH
   ½ of the training data contained unique LCSH
2. Catalogers reacted more positively than expected to results
   Interested and less afraid
3. HITL prototypes showed the most promise for future experimentation

Caroline's burning questions

1. Would faceted subject headings (post-coordinated) be more successful than subject strings, a la LCSH (pre-coordinated) in ML processes?
2. Which subject categories are more successfully cataloged using ML?
3. Could a model be trained to accurately predict LC Classification and/or Dewey Decimal Classification
4. What will the Library's policies/decisions be?
    1. Copyright concerns
    2. Accuracy vs. Relevancy
    3. Training data bias

ECD3: Extending Experiments to Explore Computational Description

1. How can ML methods support the CIP cataloging workflow?
2. How can CIP metadata generated through ML be ingested and used in BFDB
3. How can additional elements added to BF descriptions improve quality and usefulness of the metadata compared to ECD1 and ECD2?

Experiment with three different AI approaches

Data: Use data that more closely matches what catalogers work with on a daily basis

Output: Create BIBFRAME descriptions that can be loaded to test BFDB

- Require more metadata beyond the 6 fields required in task order 1

1. Allow for cataloger review in the BIBFRAME Editor
2. Extension of cataloger assisted prototypes

ECD Roadmap

ECD1

- Test multiple methods with ebook data
- Understand performance baselines
- Initial review of data quality

ECD2

- Provide more ebook data for training and tuning models
- Establish quality baselines per field
- Prototype more HITL catalog assistance workflows

ECD3

- Test methods in real cataloging workflows
- Refine and document quality of output with manual reviews
- Output data in BIBFRAME rather than MARC

ECD4 – requirements for production integration

Thanks!

Caroline Saccucci, csus@loc.gov
Abbey Potter, abpo@loc.gov