

Results of AI Experimentation for Cataloging at the Library of Congress

Caroline Saccucci & Abigail Potter
Library of Congress



Exploring Computational Description

What AI methods are most promising to make the cataloging process more efficient?

Core Principles

- Support the staff cataloger
- Maintain high quality records

Experimentation Method

- Test multiple models & approaches
- Evaluate w/ metrics & expert review
 - Focus on cataloger pain points

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

What was tested

ECD1

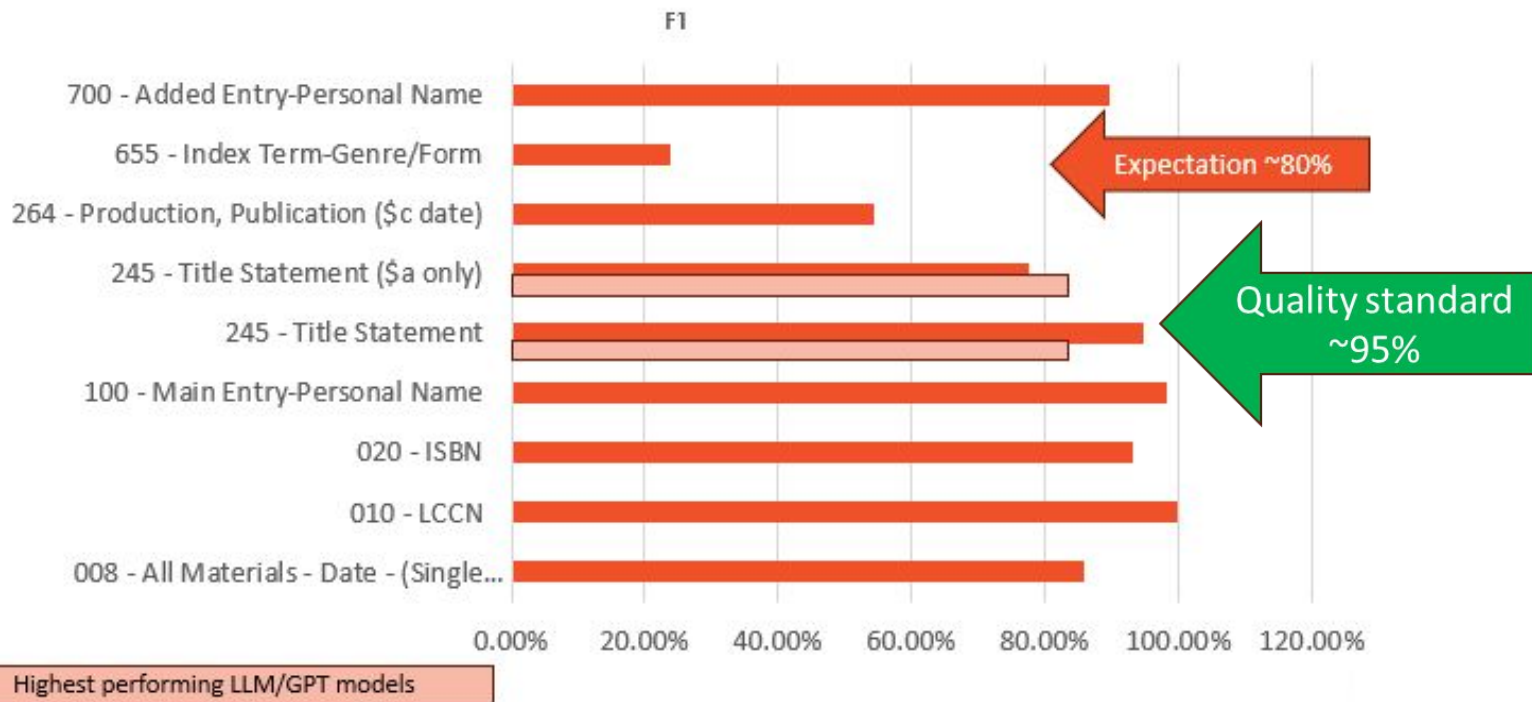
- **Models:** Bert, Spacy, GPTs with variations (NLP, NER, LLMs, transformer and non-transformer)
 - Token classification
 - Text classification
 - Data serialization
- **Human in the Loop (HITL) workflows:** combining AI output and human review or verification.

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

ECD1: Results: Text Classification, sample

Scores for token classification models by field



ECD1: Results: Text Classification, sample

Moontrap

DON BERRY

introduction by Jeff Baker

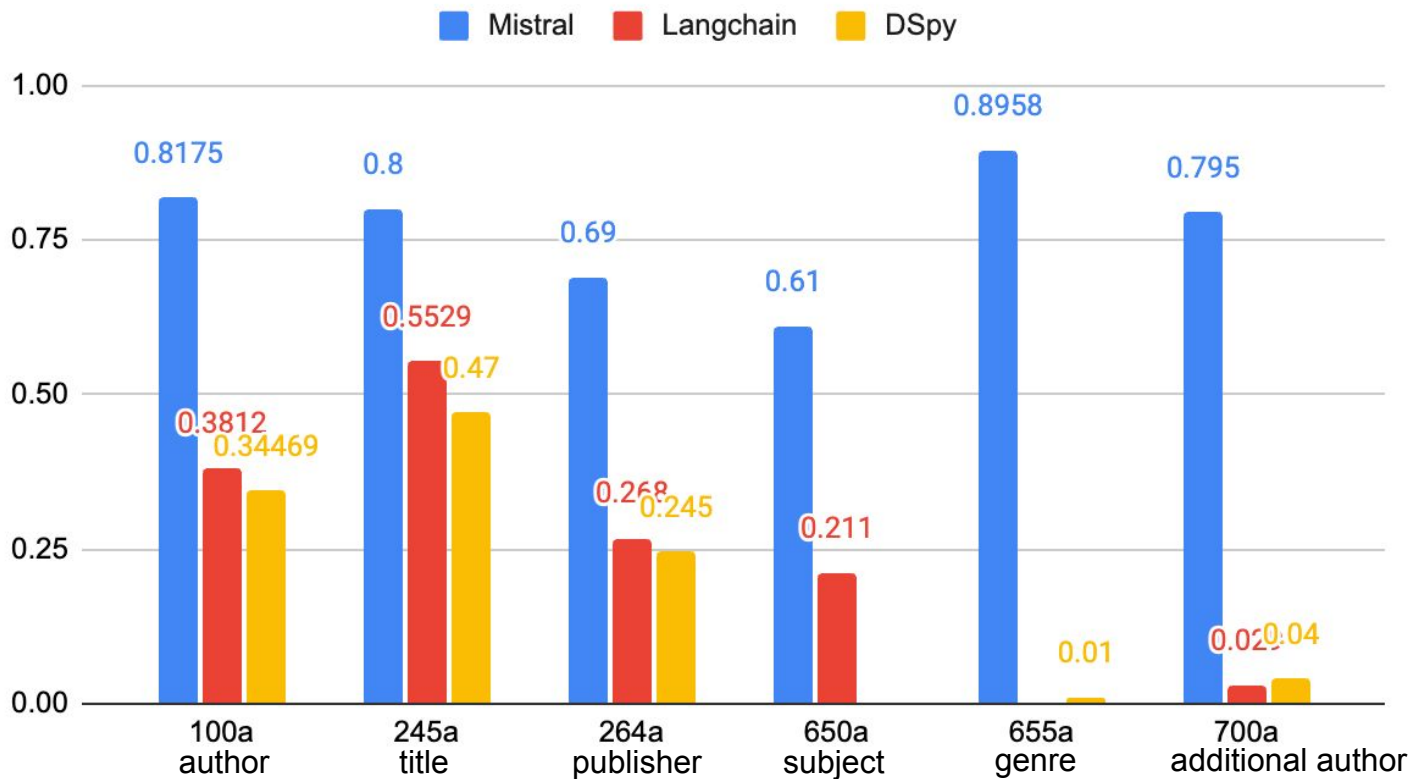
Oregon State University Press
Corvallis

Results after applying Annif

Green shaded areas are exact match to MARC XML

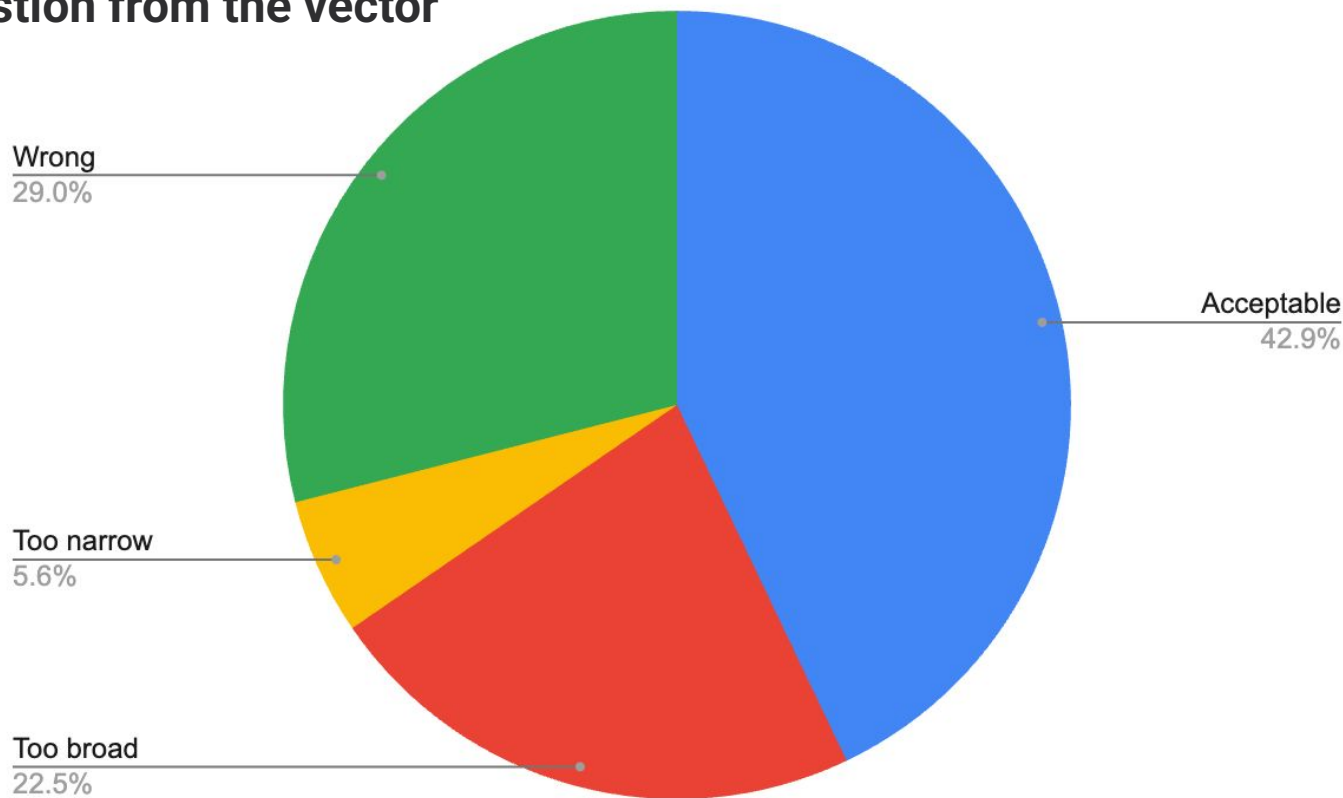
<http://id.loc.gov/authorities/subjects/sh85013380>	Berries
<http://id.loc.gov/authorities/subjects/sh2008107177>	Married people--Fiction
<http://id.loc.gov/authorities/subjects/sh97002963>	Cyberspace
<http://id.loc.gov/authorities/subjects/sh2009125638>	Fur trade--Fiction
<http://id.loc.gov/authorities/subjects/sh2008103545>	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

ECD2: Assisted Cataloging HITL Workflows

Record: 2021697918: From leadership theory to practice : a game plan for success as a leader /

[Record](#) [Subjects](#) [Summaries](#)

Title

From leadership theory to practice : a game plan for success as a leader /

Author

Robert Palestini

Date of publication

c2009

Publisher name

Rowman & Littlefield Education

ISBN

9781607090243

LCCN

HD57.7

[Image](#) [Metadata](#) [Related records](#)

< page 1 of 200 >

← back to related

aligning mind and heart : leadership and organization dynamics for advancing K-12 education / Chris Heasley, Robert Palestini

[Title page](#) [Table of contents](#) [Metadata](#) [Summaries](#) [Subjects](#)

Keywords

copy	text	score
☐	leadership behavior	6.12%
☒	human resource leadership behavior	5.97%
☐	structural leadership behavior	5.92%
☐	symbolic leadership behavior	5.85%
☐	political leadership behavior	5.84%
☒	symbolic frame leadership behavior	5.83%
☒	human resource behavior	5.83%
☒	coach	4.31%
☒	appropriate behavior	4.25%
☐	Players	4.19%

copy 5 subjects to 2021697918

Save

discard

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

Exploring Computational Description

Results

Performance of models do not meet quality standards for fully automatic operation

- **~80% accurate across all fields**
- **Up to 99% accurate on title and author fields**
- **Below 50% accurate on subject and genre fields**

Informing next steps

- Understand quality implications for BIBFRAME vs MARC (ECD3)
- Establish human in the Loop/AI in the Loop workflows, clear quality standards, and quality assurance processes
- Study the benefits and risks across workflows and user groups
- Seize the opportunity to modernize cataloging policies and practices

Ongoing 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

Thanks!

Abbey Potter, abpo@loc.gov