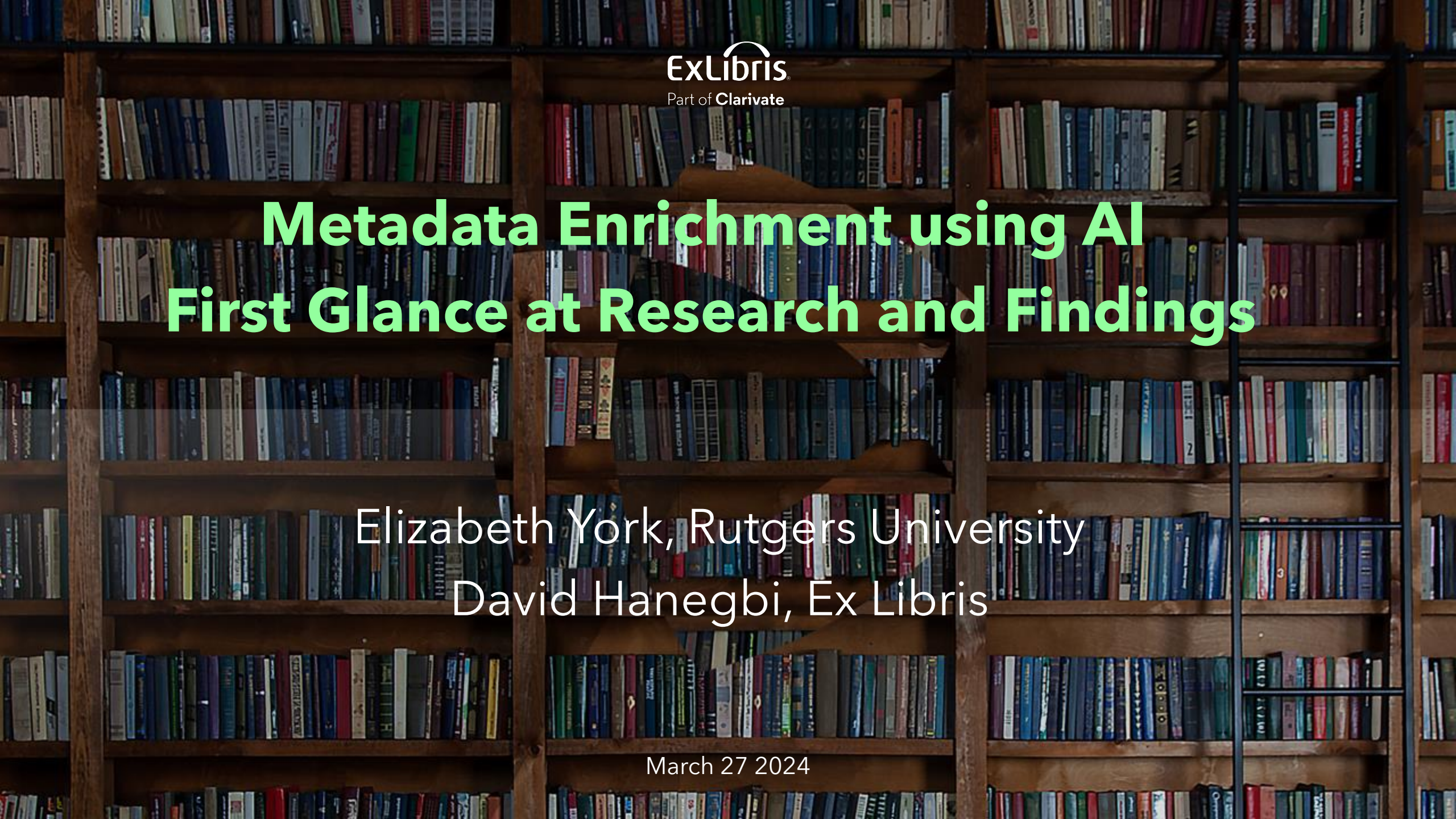


Ex Libris
Part of Clarivate

Metadata Enrichment using AI
First Glance at Research and Findings

Elizabeth York, Rutgers University
David Hanegbi, Ex Libris

March 27 2024

A Journey of a Thousand Miles Begins With a Single Step

Improving Bibliographic Records

What can AI do?

The Need

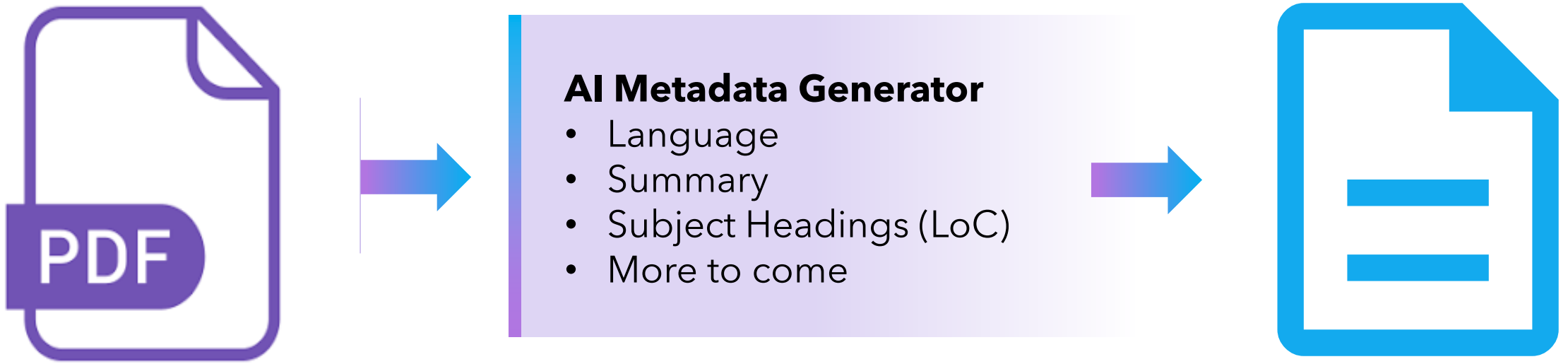
- **CZ quality – Ex Libris Focus**
- **Data Excellence** (2022 and on-going):
  - Top books collections are in high quality
  - Average score of the CZ records climbed from 72.8 to 76.1
  - 2024 goal: Books collections above 100 activations
- **How are we doing it?**
  - More sources (Providers, Library of Congress, Books in Print)
  - More tools (use of Non-MARC feeds)
- **AI** is another tool to help in this effort

The Need—Customer Perspective

- Some providers are unable to share detailed metadata with Ex Libris, so their Alma Community Zone (CZ) records have minimal metadata
- Robust metadata is needed in Primo so users can discover resources, evaluate their usefulness, and link to related resources
- Standardized metadata, such as Library of Congress Subject Headings (LCSH), Library of Congress Classification (LCC), and Dewey Decimal Classification (DDC), can be used by libraries for collections analyses in Alma and Alma Analytics
- With the public availability of generative AI, we have an opportunity to apply this new tool to our field to help fill in this missing metadata

Ex Libris AI Metadata Generator

For Alma Community Zone



Original Full Text

Enriched MARC Record

ExLibris
Part of Clarivate

AI Metadata Enriched Record

	040	__ |a UtSlPG |b eng |c IL-JeEL
Language	041	__ |a eng
	100	__ |a Scheele, Carl Wilhelm |d 1742-1786 |e Author
	245	10 |a Discovery of Oxygen, Part 2
	260	__ |b Project Gutenberg
Abstract	520	__ |a The book 'Discovery of Oxygen, Part 2' by Carl Wilhelm Scheele, delves into the discovery of oxygen and the associated experiments conducted by the author. Through a series of experiments, Scheele explores the nature of fire, air, and their compounds, challenging existing theories. He presents the idea of different types of air, each with its own properties, and highlights the need for experimental proof over mere conjecture. He also discusses his experiments with substances like sulphur and alkali, and their impact on the properties of air. This work is aimed at those interested in chemical science and the history of scientific discoveries. |7 Generated by AI.
	588	__ |a Part of the metadata in this record was created by AI, based on the text of the resource.
	542	__ |f Public domain in the USA. |k Project Gutenberg |i 2008-08-09
LC Subject Headings	650	_0 |a Chemistry |x Experiments |7 Generated by AI.
	650	_0 |a Oxygen |7 Generated by AI.
	906	__ |a BOOK

Proof of Concept and Librarian Review

- Collaboration of Clarivate's Data Science team and the Content Operations team
- We chose 1000 titles and ran AI queries on them
- Results were compared with the metadata of enriched records
- And where checked manually by librarians

- **Conclusion:** continue in first stage with Language, summary and subjects: Improve the quality and release.

Librarian Check: Timeline

- Summer 2023: Agree to participate in the project
- August 2023: Discuss MARC fields to potentially generate via AI. Ex Libris decides to focus on:
  - 041 (Language)
  - 050 (Library of Congress Classification)
  - 082 (Dewey Decimal Classification)
  - 520 (Summary)
  - 650 (Subject)
- August 2023: Check AI-generated 041, 050, 082, 520, 650 for 4 books.
  - For 650, AI generated keyword subjects, but they weren't LCSH (or any other controlled vocabulary).
  - For 050 and 082 (LCC and DDC), AI generally correctly chose the beginning part of the call number but struggled with subsequent digits.
- September 2023: Check AI-generated 041, 050, 082, 520, 650 for 15 books.

Librarian Check: Timeline

- December 2023: Ex Libris decides to focus on 041, 520, and 650. Review AI-generated 650 fields.
  - For each book, reviewed the keyword subjects chosen by AI and AI's first attempt to map them to LCSH.
  - For each book, reviewed 2 sets of subjects: one based on AI reading the whole book and one based on AI reading just the 1st 15 pages. For scholarly nonfiction works, the set from the 1st 15 pages tended to be better.
- December 2023: Review AI-generated 520 fields.
  - For each book, reviewed 2 summaries: one based on AI reading the whole book and one based on AI reading just the 1st 15 pages. For scholarly nonfiction works, the set from the 1st 15 pages tended to be better.
- January/February 2024: Ex Libris asks Community Zone Management Group (CZMG) which MARC fields to use to show a CZ record contains AI-generated metadata. We agree on 588, 035, and $7.
  - CZMG also reviews a small number of AI-generated records nearing release.
- February 2024: Librarian group checks 041, 520, and 650 on records nearing release point.
  - We checked 2 versions of each summary: a long and a short one. Sometimes, long was better because it conveyed more information, but sometimes, it was repetitive, so short was better.

Getting the Subject Headings Right

- A good recent example of LCSH chosen by AI:
  - *Exploring Religion in Our Time*: "Religion and politics," "Religion and state," "Religion and culture," "Globalization--religious aspects," and "Religious pluralism"
- For 650$a, AI needs to capture what a book is "about," (subject) not what a book "is" (genre). This is especially a challenge for fiction.
  - For *Finding Audrey*, a humorous fiction book for teens about a teen who learns to cope with mental health challenges:
    - Subject headings chosen by AI "Humor in fiction" and "Fiction" are incorrect
    - Subject headings chosen by AI "Mental Health" and "Adolescence" are good; if AI could eventually add subfield $v for form, it would be even better!
- Sometimes, AI chooses a subject that looks correct, but LCSH defines the term more narrowly:
  - For *Materiality of writing in early Mesopotamia*, a book about text-bearing artefacts from ancient societies, the AI-chosen subject heading "Societies" was incorrect because "Societies" in LCSH refers to social clubs and organizations
  - For *The philosophy of human rights*, a book that explores universal human rights, the AI-chosen subject heading "Universalism" was incorrect because in LCSH, "Universalism" refers to the religious belief that all people will attain salvation

Getting the Summary Right

- AI generally did the best job of summarizing nonfiction scholarly works, probably because the table of contents and introductory material provide a good overview of the book.
  - "Animals and Medicine: The Contribution of Animal Experiments to the Control of Disease is a book that examines the role of animals in medical research. It explores how animal experiments have contributed to the understanding and treatment of various diseases. The book covers topics such as the development of vaccines, the use of animals in organ transplantation, and the production of insulin. It also discusses the history of diseases like smallpox, rabies, and diphtheria, and how animal experiments have helped in their prevention and treatment. Overall, the book provides a comprehensive overview of the important role animals have played in advancing medical science."

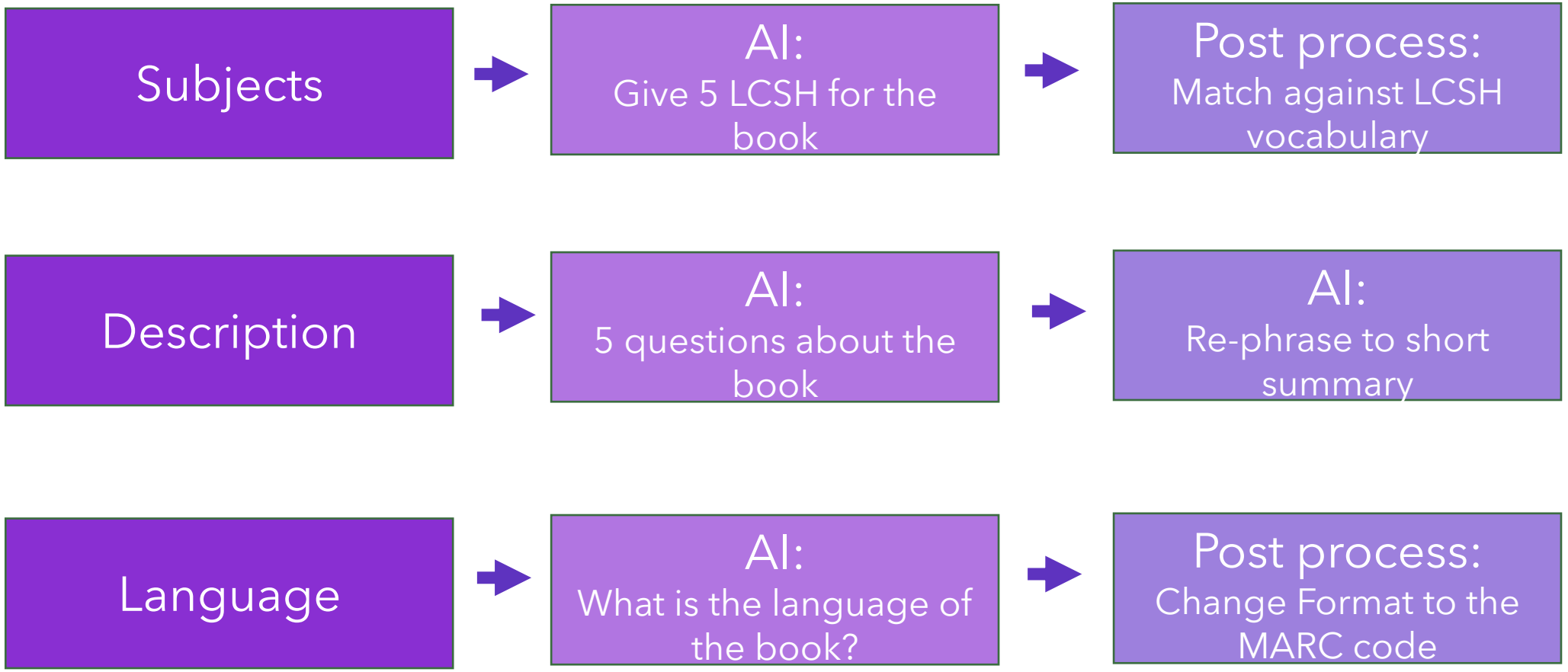
Getting the Summary Right

- AI describes many things as "fascinating" or "captivating," which can seem childish or insensitive:
  - *Trace metals and infectious diseases*: "...Get ready to dive into the fascinating world of microbes and metals!"
- AI is very positive about every book, which can seem overly promotional:
  - *Earth 2020: An Insider's Guide to a Rapidly Changing Planet*: "...It provides essential reading for anyone seeking a deeper understanding of the past, present, and future of our planet."
- AI sometimes draws too heavily on front matter, such as a series description or author note:
  - For a scholarly nonfiction work: "...A heartfelt book expressing gratitude to the people who supported and guided the author throughout their research and writing process."
- Fiction, poetry, and biographies are a challenge: Summary needs to be based on the whole book, should convey genre, and should appeal to its intended audience.

Watching for Bias in AI

- AI may replicate biases found in the books themselves.
  - When producing a summary, AI can copy dated or insensitive language found in the book.
  - It can be difficult for AI to summarize a book's claims without seeming to endorse them.
    - *21 Days to Master Numerology*: "...a self-help guide that utilizes the principles of numerology to provide readers with insights into their inner selves and life paths."
    - It is better if AI can attribute these claims to the author: "The author aims to help readers gain self-awareness, understand their life's purpose, and achieve personal growth."
- Library standards, such as LCSH, LCC, and DDC, are not without bias. AI must be trained to use library standards, but if the standards themselves are biased, AI will replicate their biases.
  - As librarians, we need to update our standards (by changing outdated terms and adding needed new terms) so systems that use our standards can do better. The changes should be made in a way machines can easily understand so we can efficiently apply them retroactively.
- As a human, when I describe resources, I take extra care with resources that document (or address topics related to) human rights abuses, genocides, atrocities, and disasters. Can AI be equally careful?

How Does it Work?



Subjects Post Processing – Examples

AI: Artists – Correspondence

Post process: $a Artists $v Correspondence

(Splits to main and sub)

AI: Professional development

Post process: Career development

(Uses internal references (450->150)

AI: Business Strategy

Post process: Business planning. Similarity score: 0.929

(Finds the most similar Subject)

The Recent Release – 200 Titles

- Ebook Central titles with no Subjects / Description
- Quality Assurance on each title, but not manual cataloging.
- Will continue with ~100 books per month

- Subjects and Description can be missing as AI didn't provide good metadata
- Subjects: In most cases, the process provided only the 650$a
- There may be important subjects which are missing
- Language: only when the current language was wrong, and AI found correct one (~5% of titles)

Future Development / Enhancements

Priority 1:

- Improve quality of Subjects, Description

Priority 2:

- **Classification** (050, 082)
- **Extract metadata from the book itself**: Subjects, TOC, Authors, publishing information etc.
- **Metadata from metadata**: can we create subjects or classification from title, Description, Table of Contents?
- Automatic Quality assurance

- We will **scale up** to more titles this year, alongside with the quality improvements