
Pragmatic AI & BIBFRAME

Recent Explorations



Area and Scope of Investigations

- Range of implementation levels from being used in production to a one off experiment
- Goal is to point out useful and practical applications and directions that intersect with BIBFRAME
- Areas we'll look at:
 - Vector Embeddings
 - Entity reconciliation
 - Work relationship classifying
 - Small machine learning model applications
 - AI features for BF adoption



“AI”

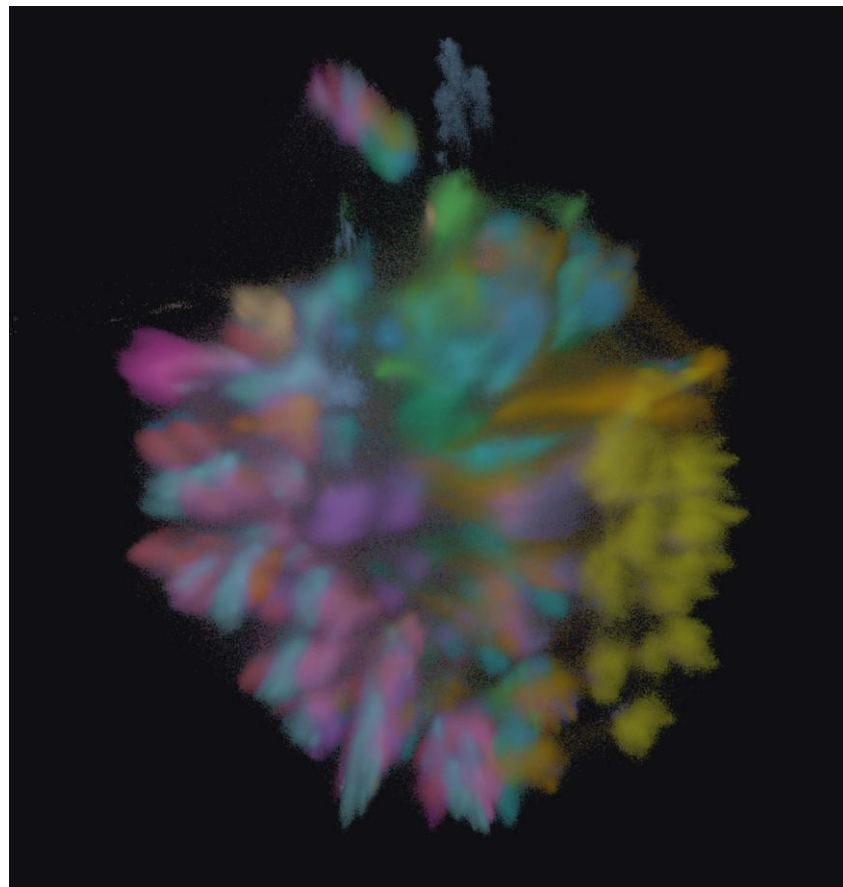
- I will try to say “Gen AI” or LLM vs “Machine Learning” to differentiate between the two.
- I feel like many of these ideas are low hanging fruit, trying to quantify, test, think of how implement.
- Shift towards thinking about how Open Weight models can be used instead of requiring frontier models.

Vector Embeddings

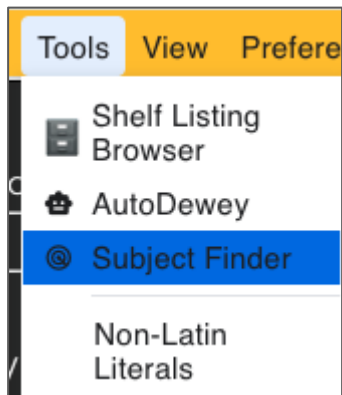
- AI Sandbox: Access to a Amazon Web Service account. Allowed to use AI services.
- Built over 20 million vector embeddings by flattening the work instance into a flat descriptive metadata text representation using the AWS titan-embed-text-v2 embedding model (1024 dimensions)
- Loaded them into a S3 Vector Bucket
- Uses Bedrock Anthropic Haiku model to do rerank and query expansion
- A reusable query service and dataset that can power many applications:
 - Subject Finder, a LCSH Suggest service in Marva
 - Discovery layer prototype
 - Dataset for analysis and visualization

Vector Embeddings - Dataset and Visualization

[https://lcnetdev.github.io/
embedded-catalog-viz/](https://lcnetdev.github.io/embedded-catalog-viz/)



Vector Embeddings - Marva's Subject Finder



Subject Finder

Close

Title: Ghosts of sicily: charles "lucky" luciano, us naval intelligence, and the cloak of secrecy that forever changed the military, the mafia, and the course of war

Summary: "It's 1942, and New York City is at war. German U-boats are sinking ships just miles offshore and Washington DC is convinced that waterfront spies are providing intelligence targeting the ships. To th...

Creator: Harmon, Mark

Contents: (not found)

Or describe:

Optional. Type a description of the work to use instead of Summary/Contents.

Get Subject Recommendations

Toggle LD Panel

Record in progress

Vector Embeddings - Marva's Subject Finder

1. Generate a flat text version of the in process record
2. Convert to embedding vector
3. Ask for the 10 nearest records
4. Open those BF records and extract LCSH headings
5. Rank which headings are most relevant to the in process records.

The screenshot displays the 'Subject Finder' interface, which is divided into two main panels: 'Subjects' and 'Classifications' (with 13 items). The 'Subjects' panel is active and shows a list of 'Recommended Subjects'. Each subject entry includes the subject name, a link to the 'LC Subject Headings' source, and the number of 'usage' records. To the right of each entry is a blue 'Insert' button. The subjects listed are:

- Espionage--United States--History (112 usage)
- Luciano, Lucky, 1897-1962 (Library of Congress Subject Headings source)
- Mafia--United States--History--20th century (38 usage)
- Organized crime--New York (State)--New York--History--20th century (25 usage)
- Spies--United States
- United States. Office of Naval Intelligence (Library of Congress Subject Headings source)
- World War, 1939-1945--Campaigns--Italy--Sicily
- World War, 1939-1945--Military intelligence--United States

The right panel shows a list of subject headings with corresponding 'Insert' or 'Insert non-LC' buttons. The headings include:

- Mafia trials--New York (State)--New York (5 usage)
- Mafia--Italy--Sicily
- Mafia--Italy--Sicily--History--20th century (65 usage)
- Mafia--United States
- Organized crime--New York (State)--New York
- Pearl Harbor (Hawaii), Attack on, 1941 (214 usage)
- Pearl Harbor (Oahu, Hawaii), Attack on, 1941 (Insert non-LC)
- Pearl Harbor, Attaque sur, 1941 (Répertoire de vedettes-matière source) (Insert non-LC)
- Secret service (Faceted application of subject terminology source) (Insert non-LC)

At the bottom of the interface, there are buttons for 'More Suggestions', 'Search Again', and 'Done'. A 'Toggle LD Panel' button is located at the bottom left.

Subject and Classification Recommendations

Vector Embeddings - Marva's Subject Finder

- Tested in Marva with a dozen catalogers for a few weeks. They reported:
 - Mimics existing partial copycat workflows, just does it faster.
 - Useful when working in a unfamiliar domain or to explore possibilities.
 - Not useful when working in well know domain.
- In production, any Marva cataloger can use it with supervisor's permission

Vector Embeddings - Discovery

1. Take query and use LLM to generate a “perfect” resource description for that topic, basically hallucinate a record.
2. Vectorize that phantom resource
3. Find the resources that would be its neighbors, those are the most relevant records to the question

Internal proof of concept demo

Vector Embeddings - Takeaways

- Bibliographic records are well suited to RAG methodology
- All services are language/script agnostic.
- Could be a tool to enable a more relationship based discovery system (GraphRAG).
- Built in guardrails against LLM shenanigans.
- Extremely inexpensive to create and run.

Entity Reconciliation

- BIBFRAME is a shift towards Entities and Relationships. Ensuring they are correct as well as adding and enriching them is important to achieve a network effect.
- A couple examples from the Wikidata/EMCO world.
 - Instead of trying to get a LLM to catalog a book how about we just ask it to help find mistakes? Applying LLM reasoning to QA checks.
 - How well does it perform at reconciliation tasks at scale?

EMCO LCCNBot Check

- When a NACO cataloger adds a reference to Wiki* it will pick up the reference and insert the NAR's LCCN into Wikidata. Fully human generated matching.
- Over the last year I have had a LLM check process that doesn't prevent the match from being made but flags it on a QA report.
- <https://id.loc.gov/loads/lccnbot/>



False

▼ info

Despite sharing the name "Zhang Ping" and birth year 1970, the Chinese characters for "Ping" are different (平 vs 苹), and their professional fields (mathematics vs. history) and affiliations are entirely distinct, indicating they are different individuals.



False

▼ info

The source entity, Alexias Hatun, was a historical figure who died in 1488, making it impossible for her to be the translator associated with publications from 2023 and 2025 as described in the possible match.

Bulk Enrichment Test

- Historically NARs would have citations like "Wikipedia, viewed Mar. 4, 2011" but no other information. 65,000 NARs exist with no LCCN link to Wikidata. What percentage of the 65,000 could be automatically linked via LLM?
 - Search the name in Wikidata
 - Compare each Wikidata result to their MADS counterpart document
 - Judge if they are the same entity
 - I used the open weight model Qwen3.8 27B 8-bit
- Sampling I found there is a 1% error rate over all
- As scenarios became more complicated the error rate grew:
 - Pseudonym - 1.5% error rate
 - Name merges (schools changing names for example) - 17% error rate

Entity Reconciliation - Takeaways

- It demonstrably works, even open weight models have ability to help in entity reconciliation.
- These examples demonstrate reconciling to external data but next area of testing will be to aid during internal reconciliation workflows.

Relationship Classifying / Cluster Building

- Relationships between other works and instances are very easy to communicate in BIBFRAME but there is a lack of them in historical data.
- Only about 20% of the catalog would even have potential relationships.
- Debatable usefulness for use while cataloging but could be very helpful for discovery. Tests framed in the idea of building super work clusters for discovery.
- Steps
 - Cast a very wide clustering net based on title/contributor
 - Ask the LLM to generate relationships between all the clustered works
 - Used open weight model Qwen3.6-35B-A3B-bf16

Relationship Classifying / Cluster Building

- Results:
 - 97% correct with coarse relationship
 - 87% correct with fine grain relationship
- [Read more](#)
- [Cluster Viewer](#)
- Needs more work. Most useful for discovery layers, enriching the graph.

Search: on the road Random

book · 10 associated resources · cluster of LC records sharing one super work



On the road
Kerouac, Jack, 1922-1969 · 1957-2016

Jack Kerouac's novel 'On the road' and its various editions, formats, adaptations, and derivative works (including the original scroll, abridged audio, and musical compositions inspired by it).

10 RESOURCES 3 FORMATS 2 LANGUAGES 1957-2016 SPAN

Editions & Printings

superswork_swd BOOK

On the road ★

Kerouac, Jack, 1922-1969 · New York : The Viking Press, 1957. · 310 pages

Title 'On the road', The Viking Press, 1957, note 'First edition'; the original publication of the novel.

later_printing_reprint BOOK

On the road

Kerouac, Jack, 1922-1969. · Cutchogue, N.Y. : Buccaneer Books, c1975. · 310 p.

Title 'On the road', Buccaneer Books c1975; reprint of the novel.

critical_or_annotated_edition BOOK

On the road

Kerouac, Jack, 1922-1969. · New York : Penguin Books, 1979. · vii, 605 p.

Title 'On the road', text and criticism edited by Scott Olson; published by Penguin 1979 as part of The Viking critical library; includes scholarly apparatus.

abridged_edition AUDIO (SPOKEN) AUDIO

On the road

Kerouac, Jack, 1922-1969. · Saint Paul, MN : Penguin Highbridge Audio, 1986, c1993. · 2 sound cassettes (156 min.)

Title 'On the road [sound recording]; note 'Abridged'; read by David Caradine, published by Penguin Highbridge Audio.

later_printing_reprint BOOK

On the road

Kerouac, Jack, 1922-1969. · New York, N.Y. U.S.A. : Penguin Books, 1991. · xxiil, 310 p.

Title 'On the road', Penguin Books 1991, part of Penguin twentieth-century classics; standard reprint of the novel with an introduction by Ken Chertner.

later_printing_reprint BOOK

On the road

Kerouac, Jack, 1922-1969. · New York : Viking, 1997. · 307 p.

Title 'On the road', Viking 1997; standard reprint of the novel.

revised_edition BOOK

On the road : the original scroll

Kerouac, Jack, 1922-1969. · New York : Viking, 2007. · 408 p.

Title 'On the road : the original scroll', edited by Howard Cunnell, published 2007; this is a distinct textual realization of the work based on the original manuscript scroll, differing from the standard published text.

later_printing_reprint BOOK

On the road

Kerouac, Jack, 1922-1969. · New York, N.Y. : Viking, 2007. · 307 p.

Title 'On the road', 50th anniversary ed.; published by Viking 2007; same author; standard text; commemorative printing of the original work.

later_printing_reprint BOOK

On the road

Kerouac, Jack, 1922-1969. · New York : Penguin Books, 2016. · 307 pages

Title 'On the road', Penguin Books 2016, part of Penguin orange collections; standard reprint of the novel.

EDITION TIMELINE - 1957-2016

- 1957★ The Vik...
- 1975 Buccane...
- 1979 Pengi...
- 1986 Pengi...
- 1991 Pengi...
- 1997 Viking
- 2007 Viking
- 2007 Viking
- 2016 Pengi...

Derivative & Related Works

derivative_work SCORE

Small Machine Learning Applications

- We have a lot of data to train small task specific models with to solve problems that are too complicated for a rule based or programmatic approach.
- Example:
 - Exploring how we would do authority work in MADS in Marva. Requirement to convert to MARC for distribution.
 - We want to move away from subfield encoding in literals that MARC depends on.
 - Trained a Conditional Random Field (CRF) model on NAR data to generate a MARC Key subfield string that can be used in the conversion but the cataloger does not need to subfield the string.
 - 97.5% successful based on random sample
 - Demo: <https://thisismattmiller.github.io/naf-crf/>

Marva Scan - OCR / Structured Data Extraction

- Connects Marva to phone or web camera to send images to vision supported LLM to extract structured data to insert into the record.

The screenshot displays the Marva Scan interface, which is a web-based tool for processing scanned documents. The main window is titled "Marva Scan • Title page processing..." and features a "Scan" tab. A message at the top states: "No LCCN or ISBN found in record — will use Marva E number instead (likely want to add LCCN or ISBN)". Below this, there are options for "Phone" and "Webcam" (selected). A video feed shows a hand holding a document. To the right, a section titled "Existing data" states: "No ISBN or LCCN found on this profile to look up." Below the video feed, there are buttons for "Cover", "Summary", "Title page" (selected), and "Copyright". Further down, there are buttons for "Just OCR", "Non-Latin OCR", "TOC", and "Back Cover". A "Connection" field shows a URL: "dQVCj5cF-TAHCLEfg=e1778594873338". At the bottom, there is a section titled "Insert from scanned data" with a "Summary" button and a "Live scan" button. A "Text" field contains the following text: "With more than ten thousand entries, The Book of Positive Quotations is an invaluable tool for writers, public speakers, coaches, business leaders—anyone who needs to communicate a positive message to an audience. It is conveniently organized by subject and thoroughly indexed for ease of use." The background shows a sidebar with various categories like "Creator of Work", "Title information", "Contributors", "Notes about the Work", "Contents", "Subjects", "Geographic coverage", "Genre/Form", and "Classification numbers". The top navigation bar includes "Menu", "Tools", "View", "Preferences", "Layouts", "Save", "Validate", "Post", "Open in LCAP", "STAGING: Monograph", "Record Info", "DCTap: marva-prod", and a user profile "mattmiller (mm1234)".

Marva Scan - OCR / Structured Data Extraction

Menu Tools View Preferences Layouts Save Validate Post Open in LCAP STAGING: Monograph Record info DCTap: marva-prod mattmiller (mm1234)

Expression of [search for a] (expression or position for a field)

Close

Delete Instance? 3.2.0 3.1.0 3.2.0 cam a2 001 inB 003 DLC 005 202 007 t| 008 260 000 e c 040 \$a DLC \$b eng

Insert from scanned data

Title live scan Insert into Instance and Work

Main title The Book of Positive Quotations

Subtitle (optional)

Statement of responsibility live scan Insert into Instance

Statement Compiled by John Cook ; edited by Steve Deger & Leslie Ann Gibson

Edition statement live scan Insert into Instance

Statement This Lyons Press paperback edition: 2024

Publication info live scan Insert into Instance

Date 2024 Publisher Lyons Press

Place code City

ctu Essex

→ Connecticut

Primary author live scan Insert into Work

Name Cook, John

Role(s) compiler (com)

Contributor live scan Insert into Work

Name

David Stein

Note

Terms of availability

Marva Scan

Scan Images Debug

Scan with your phone

Ready — scan the QR code with your phone.



Existing data

66082276 no data

Connection Resource ID c_1xBdQvIAMCLDg= 66082276

Open mobile link

Reconnect

Insert from scanned data

Capture a title page, summary, back cover, table of contents, or OCR to enable insertions.

Marva Scan - OCR / Structured Data Extraction

- Used in test environment with catalogers to evaluate how well it works.
- Sends images to a Sonnet 4.6 LLM with specific prompts on how to extract data for each page
- Positive feedback, surprising useful for non-latin scripts

Questions?

Thanks for Listening!

mattmiller@loc.gov

