AI for World Bank Group (WBG) Library Intelligent Search: An Exploration of Three Technologies

Flor Trillo, Yasemin Palta, Oleksandra Postavnicha, Hamed Honari, Kutemba Chikumbi, and Kemal Olguin, World Bank Group, 1818 H St NW, Washington, DC 20433, United States

Background

The WBG Library (ITSLP) is seeking innovative solutions to develop a Discovery Tool that allows an efficient search process by providing applicable and relevant search results to library clients from numerous data sources. While some sources are freely available and easy to access for resolving results, others only provide relatively basic metadata due to the API request and require a subscription. Therefore, a tool configured to search effectively, returning and surfacing relevant and democratized results leveraging the limited metadata fields (keywords, title, source title, date, etc.) offered in most data sources are highly needed.

Considering the above challenges, the Technology & Innovation Lab (ITSTI) team of the WBG explored four different tools with key data sources to guide the design of possible solutions under the following conditions: keeping small scope in terms of data sources to work efficiently during the exploration and focusing on high-value data sources with minimal and/or varied metadata, leverage subject matter expert's feedback.

Objective

Exploring AI technologies to provide efficient support in terms of discoverability and democratized results from data sources to WBG library clients at the point of need, with a high relevance level only using metadata from the scholarly and scientific ecosystem.

Methods

Project Approach

After doing research, including interviews, deep investigations, and identifying key pain points to focus on, which oriented to the search process across various data sources with limited metadata. The team collectively decided on several databases that are both high value and/or could pose a problem in resolving good results to select samples (18,411 records in total) of three different key data sources (DS).

1. **Set the challenge space and scope for Prototyping**
2. **Get up to speed on the context of the challenge and data sources**
3. **Build out Prototype leveraging dataset within the scope**
4. **Test for accurate and desirable results as a result of the search**

Data Extraction

- **Data Points:** Crossref (DS-A), with 38 metadata fields in each record; Core Free (DS-B), with 31 metadata fields in each form; and our largest and internal institutional repository called, Documents & Reports (DS-C), with 46 metadata fields.

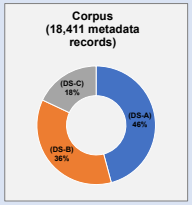
Data Extraction (continuation)

- **Queries:** Top 10 queries (July 2022) from WBG Library clients for each tool and presented results for evaluation and relevancy check.
- **Metadata:** Minimum requirements such as Title, Author, Subject, Abstract/Description, Creation Date, Language, Format, Identifiers DOI/ISSN, Source Journal, URL, and Full text.
- **Tools Data Collection:** Python to extract the top 100 items from each source.

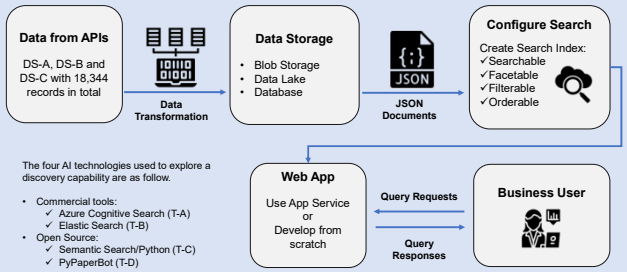
Metadata Exploration

To improve metadata exploration and ensure consistency across all the data sources, we recommend the following steps:

- ✓ Use filters and advanced search (filters could include additional fields such as creating data, file type, location, organization, data source, language, etc.);
- ✓ Generate downloadable citations, hyperlinking to external resources, use ISBNs to fetch book cover images, labeled display alternatives to ISBD;
- ✓ Fetch table of contents, redirect automatic search;
- ✓ Resource type diversification, functional diversification, and source diversification.



Technical Approach: End-to-end workflow



Results – Key findings

After running the following queries using T-A, T-B, and T-C under two different configurations: a) only titles, and b) all fields.

- ✓ Query 1 (Q1): **New York Times**
- ✓ Query 2 (Q2): **EconLit**
- ✓ Query 3 (Q3): **use of remittance**
- ✓ Query 4 (Q4): **digital currenc* or crypto* or digital money or digital asset***

We exclude T-D from the following findings due to not customizability of input data sources and the ability to only perform one query at a time.

- Q1 & Q2. We found that when shorter queries, using T-A and T-B, relevant content in the Top 15 surfaced, including all DS content when information was available. But in the case of T-C, the top relevant contents retrieved belong to one data source (mainly DS-A).
- In the case of Q1 (a newspaper) and Q2 (a database), DS-A offers the link directly to the source. DS-B complements samples of older notes for Q1 and offers more information on how to use it, basic search, and subject descriptors for Q2.
- In cases of largest queries such as Q3 and Q4, using T-A and T-B, offer results in a federated manner in both configurations, discovering content in all DS with the best results in the Top 15.
- DS-C has the largest coverage in terms of metadata fields. For that, it gets a better position when the configuration covers all fields, but in shorter queries is not always surfaced.

Conclusions and recommendations

- ✓ Consider connecting data from additional data sources using native APIs and consolidate it in a single placeholder such as Blob Storage.
- ✓ Have metadata fields to be consistent across multiple DS for the proper search index ingestion.
- ✓ To ensure a federated/better representation of metadata contained in the search tool that is not biased towards any data point sources, it is recommended to be blind to the source of the data point (i.e., DS-A, DS-B, etc.). Thus, the search results are merely based on the related areas of interest.
- ✓ It is recommended to specify if the search should be based on the field of interest (i.e., "title" only, "abstract" only, or a combination of fields of interest).
- ✓ While configuring the search index, including all relevant metadata fields to be searchable is important.
- ✓ The WBG Library will review all these options to start developing a suitable solution.

References

- API Crossref: https://api.crossref.org/swagger-ui/index.html
- API Core Free: https://core.ac.uk/services/api
- API WBG D&R: https://documents.worldbank.org/en/publication/documents-reports/api#:~:text=The%20World%20Bank%20Documents%20%26%20Report,Reports%20and%20the%20World%20Bank.
- Azure Cognitive Search: https://azure.microsoft.com/en-ca/products/search/
- Semantic Search: https://opensemanticsearch.org/
- Elastic Search: https://www.elastic.co/

Disclaimer Statement