ANALYZING

# OCLC DATA

FROM
ANNA'S ARCHIVE

FINAL PROJECT
PRESENTATION

12.13.2023
PRATT INSTITUTE

# PROJECT OVERVIEW

## Downloading Data

- The dataset was time consuming to download but made easier by the fact that our professor generously hosted the data for me to download.
- With many millions of records, the files are very large. The largest file is over 1 terabyte. I luckily have an external hard drive that has capacity to store files this large, so I had to download and store the data there. I was only able to access the data at home for this reason because my drive isn't portable.

## Assessing Records

- The largest file, title_json.jsonl, contains **over 770 million json objects**. The bulk of this project has consisted in getting to know the datasets that I chose to work with: title_json.jsonl, briefrecords_json.jsonl, and providersearchrequest_json.jsonl
- I did this by probing the data in different ways, first **in the command line** and later **using a script to sample randomly from each file** to create smaller datasets to **visualize using pandas dataframes**. This allowed me to get a sense of the metadata structure of each file.

## Sharing Findings

- The final leg of the project will consist of:
- Finishing data visualizations using pandas dataframes
- Finalizing a jupyter notebook that adequately guides a stranger through the material using explanatory markdown sections
- Creating either a lightweight website or documentation to make a github repo user-friendly to submit to the AA contest.

# WHAT IS ANNA'S ARCHIVE?

## Anna's Archive

Open source search engine providing access to unified list of shadow libraries across the web, including **25,116,839 books and 99,425,860 papers**, all mirrored on IPFS

## Why this data?

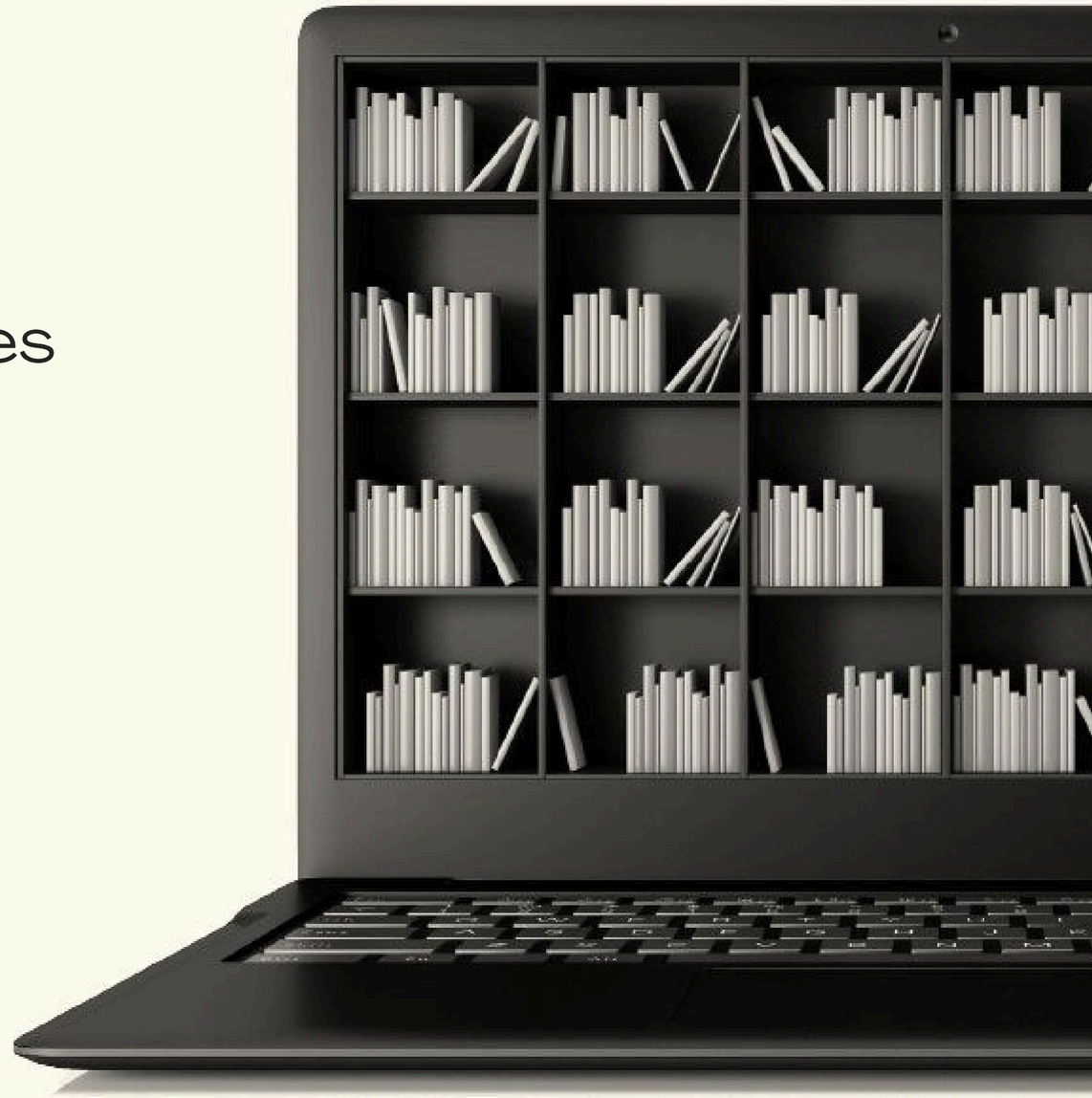
This metadata represents years of collective labor of thousands of information professionals and could greatly enhance visibility of materials in shadow libraries.

## Shadow Libraries

Shadow libraries like Library Genesis and Sci-hub have grown exponentially in past decades. These repositories provide access to material otherwise unavailable to many, in defiance of copyright law.

## How this data?

The WorldCat data was hacked from OCLC by people associated with Anna's Archive.



# THE VALUE OF METADATA

# WHAT ABOUT WORLDCAT?

Union catalog covering a large portion of the world's library collections

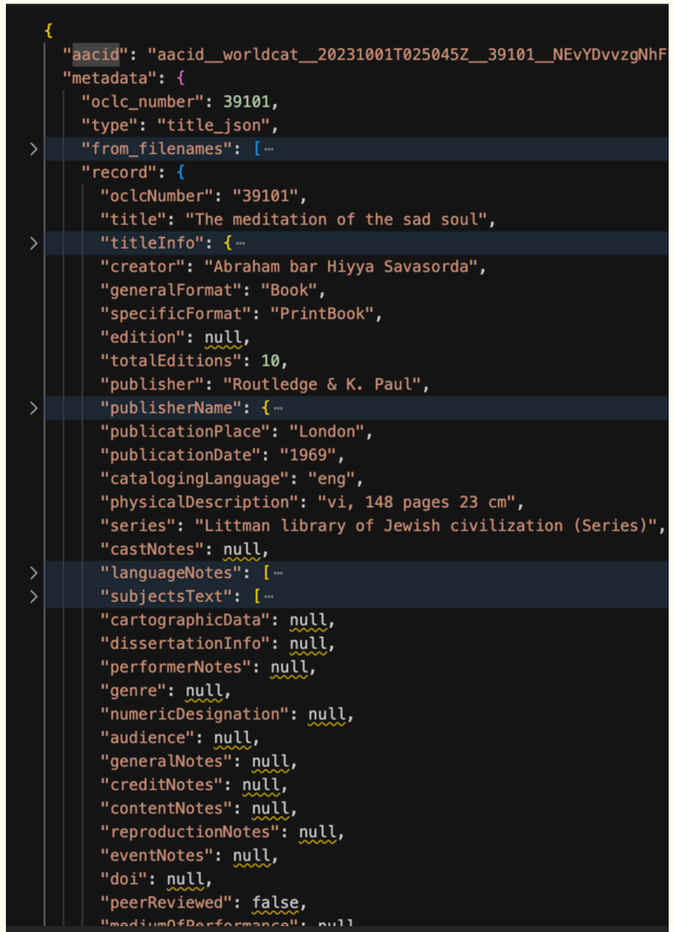
Catalog search is free,  but entire dataset is custodied by OCLC, which uses the catalog's data to power many of its paid services

According to OCLC's site, WordCat contains records for more than **3.3 billion** resources, including "books, e-books, movies, music, art, digital materials, and real-life artifacts"

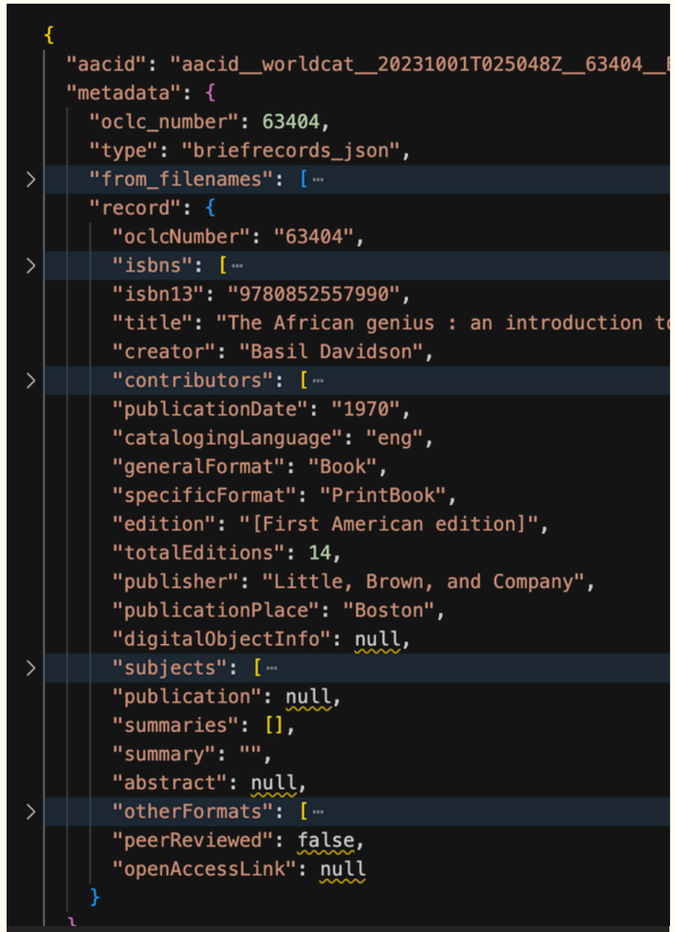
More than **500 million** of these are bibliographic records

WORLD RECORDS

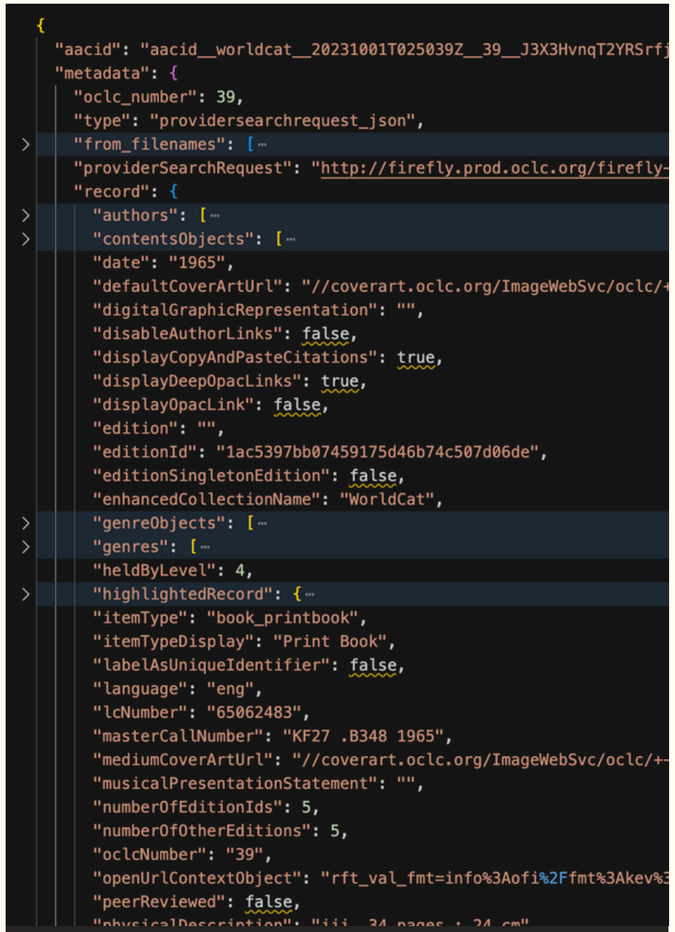
# SAMPLE RECORDS




title




briefrecords




providersearchrequest

# FINAL STEPS

The final steps of this project involve creating a lightweight website or documentation for my github repo to host the data and analysis I have performed over the course of the project. I plan to submit it to the Anna's Archive data analysis competition.

My goal in the final visualizations and investigation will be to determine completeness of the different metadata fields as well as the uniqueness of values for each key. Hopefully this will illuminate which fields might be most useful for search.

# THANK YOU FOR WATCHING