Collections as Platform: Components in a document collection case study

Based on processes developed for the Know Systemic Racism project

COLLECTION AS PROCESS

) COLLECTION AS OBJECTS

If the dataset is incomplete,

return to source to gather

more data (steps 1-4)

COLLECTION AS DATA

CROSS-COLLECTION DISCOVERY

Retrieve

Crawl URLS and return files

Gather

Create per-site folder and generate unique filenames

Provide Access

Create Globus endpoint for access for subsequent processing steps.

Assess

Load files from Globus into

Datashare for exploratory

+ Apply tags and filters

filtering, browsing and reading

Convert to text

Load files into Google Drive and Jupyter notebook environment

Extract data

Extract metadata from text files: Creation date, author, location, etc.

Refine Set

Compare tabular list of files with source URL list to assess quality of scrape and identify missing documents

Segment text

Split text into consistent record-level entities.

- + Align metadata with new structure
- + Add PURL to each record

Graph Discovery

+ Prepare documentation

discovery via Redivis

+ Configure access

+ DUA

Prepare the dataset for access and

Link/Enrich

Enrich both collection level and document-level metadata with linked authority records and knowledge bases where possible and geospatial data.

- + Produce file-level metadata.
- + Produce entity-level metadata.

Deposit zipped files into

SDR Dark

+ Minimal collectionlevel description for deposit

Gathered Files

A record of the initial collection stage for internal library use.

Location: **SDR**

Format: **Zipped bundle**

Entity/Granularity: Collection

Access: Restricted/Dark

Related files:

- + Readme
- + Checksum

SDR Discovery

Prepare the document collection for access and discovery via Searchworks

- + Replayable spreadsheet
- + Generate PURLS
- + Define APO

Document Collection

Public access to read documents in original presented format.

Location: **SDR**

Format: **PDFs**

Entity/Granularity: **Document**

Access: Public/Licensed

Related files:

- + Documentation
- + Tabular metadata

Stanford Data Explorer

Prepare the dataset for access and discovery via Redivis

- + Prepare documentation
- + Configure access
- + DUA

Tabular Dataset

Normalized dataset with consistent record-level entities for exploration, access and data analysis

Location: Redivis

Format: **CSV**

Entity: Structural element

Access: Public/Licensed

Related files:

- + Documentation
- + Codebook
- + DUA

Knowledge Graph

Entity level relationship graph for complex queries with links out to documents, knowledge bases, other relevant sources and resources. Location:

Format: Graph database (Neo4J)

Entity/Granularity: Key-value

Access: Public/Licensed

Related files:

- + Documentation