Auctus Dataset Search

Auctus is a dataset search engine that supports data discovery, exploration, and augmentation. It automatically ingests datasets from multiple sources, including public open data repositories, and provides advanced search capabilities over these datasets.

Data Ingestion

Auctus automatically (and periodically) ingests datasets from a variety of repositories and Web sites. Users can also upload datasets and provide custom plugins to ingest data from new sources.

The public Auctus instance currently indexes over 20,000 datasets

- Socrata: 18,015 (46 different domains including cityofnewyork.us, medicare.gov, sfgov.org, novascotia.ca)
- Zenodo "covid": 1,040 (datasets matching the query term "covid")
- Indicators from University of Arizona: 1,094
- Indicators from World Bank: 20
- Direct upload: 86

Supported File Formats

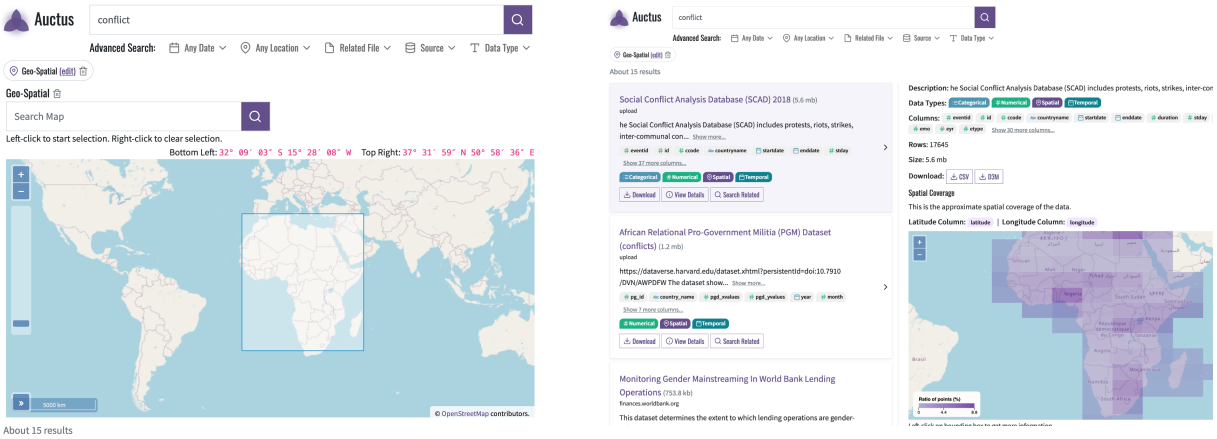
- CSV
- TSV (and other separators)
- Excel 97 (.xls) and 2003+ (.xlsx)
- SPSS files (.sav)
- Stata 114 and Stata 118 files (.dta)

Data Transformation, Profiling and Indexing

Auctus automatically transforms datasets to make them easier to manipulate and consume. For example, it:

- Transforms datasets into standard CSV
- Pivots data with dates for columns
- Automatically detects and skip "header rows" that are commonly placed at the top of an Excel files

Datasets are automatically profiled to detect the types of their (categorical, numerical, spatial, and temporal attributes), compute statistical summaries (e.g., frequent items, mean, and variance), and to derive sketches. The profiler outputs meta-data that is used to enable data discovery queries.



Searching for data about "conflict" in Africa (left). Matching datasets and their summaries can be visually explored (right).

Data Search and Discovery Queries

Auctus supports a rich set of data discovery queries: besides keyword search, users can explore dataset collections based on space, time, and data relationships. Users can also pose data augmentation queries, in which given an input dataset D, it returns a ranked set of datasets D that can be joined with or concatenated to D. Auctus also offers a user-friendly interface for search results. Filters and data exploration capabilities help users select the most suitable datasets for their information needs.

System

Auctus instances can be deployed for different domains and data collections. You can find a sample deployment at: auctus.vida-nyu.org

Besides the Web-based interface, users can interact with the search engine programmatically through Python and REST APIs.

Repository: gitlab.com/ViDA-NYU/auctus/auctus
Demonstration Video: youtube.com/watch?v=lZQbh3ctq6Q

Profiler library available on PyPI and usable from the command line (python -m datamart_profiler my_file.csv)

Materialization library available on PyPI (handles getting datasets from metadata, converting file formats, and applying autodetected fixes like pivoting and removing header rows)

Administrative areas database available on PyPI (resolve names into admin areas, identifying level, containing area, bounding box, using data from GeoNames, WikiData, and OpenStreetMap)