Datamart

by Information Sciences Institute

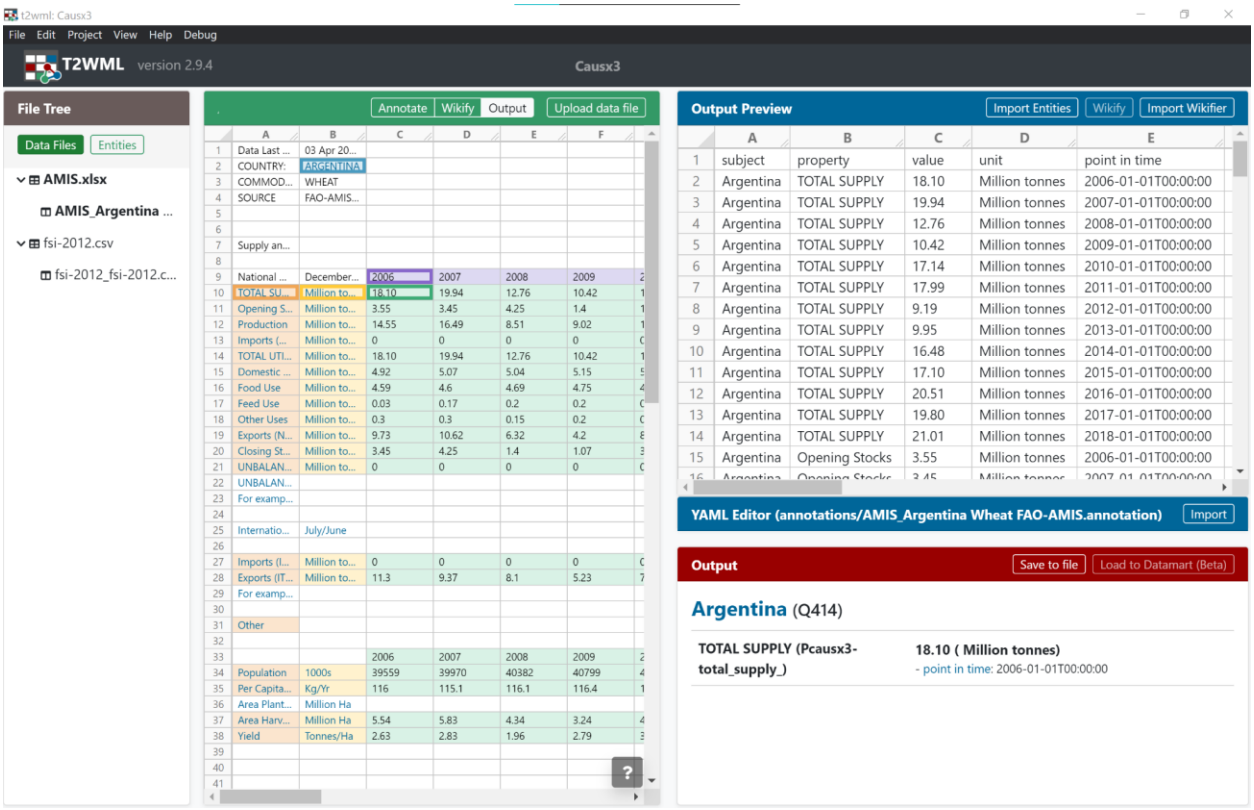
The ISI Datamart project is building technology to create the largest publicly available knowledge graph to power data-driven models in a wide variety of domains. At the core of Datamart is Wikidata, a publicly available knowledge graph that already contains over 93 million entities. Datamart will enable communities of interest to build satellite knowledge graphs that contain detailed knowledge in domains of interest. The enabler technologies includes an architecture for combining public data with private data kept within an organization's firewall, and tools to automate the ETL process required to syntactically and semantically align the data in millions of spreadsheets and CSV files to the Wikidata semantic representation.

The ISI Datamart project webpage is here https://usc-isi-i2.github.io/datamart/.

Semi-Automated Data Curation

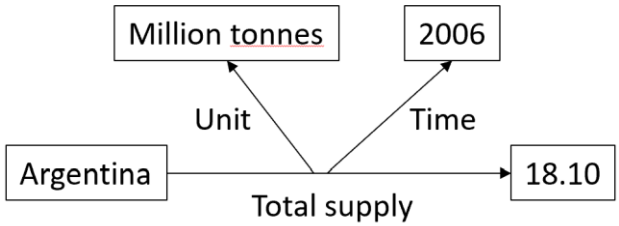
Datamart provides a semi-automated tool T2WML (Table to Wikidata Mapping Language) to transform structured data to canonical tabular format.

- T2WML automatically detects syntactic and semantic regions of spreadsheets and CSV files. The regions include main subject, property, value, unit and time. See the color coded cells in the middle pane below. The user can adjust the regions as needed.
- T2WML transforms these annotated regions into canonical format where each row is a single record, see the Output Preview pane on the right.
- The ingested data is aligned with Wikidata to enable joins across multiple data sources. Notice in the Output pane in the lower right should that the string ARGENTINA is wikified and mapped on the Wikidata entity Q414.



Datamart Core

Internally Datamart stores datasets as part of the Wikidata knowledge graph. Statements extracted from tabular data are stored as edges, e.g. the `Total supply` for `Argentia` is `18.10`. Additional context about statements are stored as qualifiers on the edges, e.g. the `Unit` is `Million tonnes` and the `Time` is `2006`.



The backend storage engine for Datamart is Postgres. The Datamart database schema is designed to be compatible with SPARQL Blazegraph and the Knowledge Graph Toolkit (KGTK).

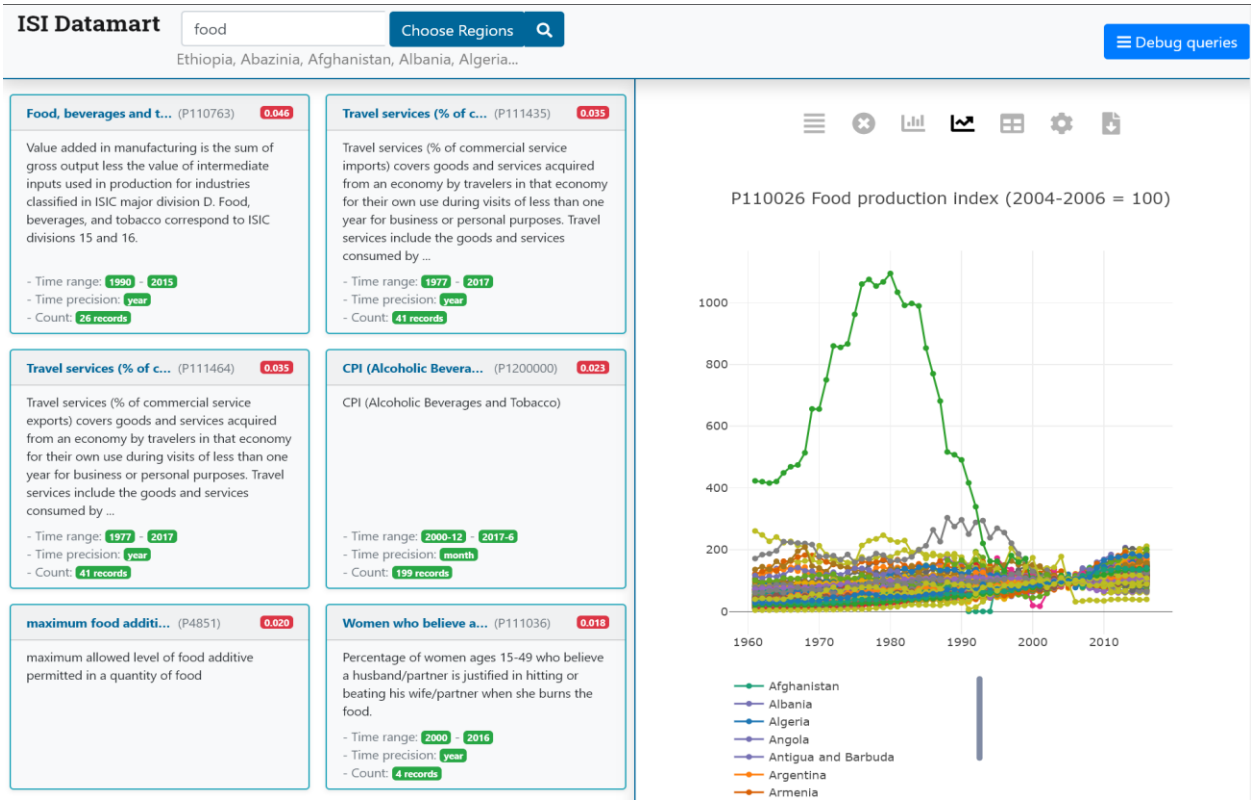
Datamart always outputs structured data in canonical format with one record per row, see below. The canonical data format is defined here.

	A	B	C	D	E
1	subject	property	value	unit	point in time
2	ARGENTINA	TOTAL SUPPLY	18.1	Million tonnes	2006
3	ARGENTINA	TOTAL SUPPLY	19.94	Million tonnes	2007
4	ARGENTINA	TOTAL SUPPLY	12.76	Million tonnes	2008
5	ARGENTINA	TOTAL SUPPLY	10.42	Million tonnes	2009
6	ARGENTINA	TOTAL SUPPLY	17.14	Million tonnes	2010
7	ARGENTINA	TOTAL SUPPLY	17.99	Million tonnes	2011
8	ARGENTINA	TOTAL SUPPLY	9.19	Million tonnes	2012

Datamart Fuzzy Search

ISI Datamart provides a graphical user interface to search for datasets. The functionality include:

- Approximate string matches of query terms to dataset metadata description
- Geographical search of countries and administrative regions.
- Search results returned as scrollable tiles
- Time series overview of datasets



System

Datamart is dockerized so that it can be easily deployed. A sample Datamart deployment is here: https://dsbox02.isi.edu:8888/datamart-api-dev/

The ISI Datamart dataset metadata schema and data schema are defined here: https://datamart-upload.readthedocs.io/en/latest/

The ISI Datamart API is defined here: https://datamart-upload.readthedocs.io/en/latest/api/. This Jupyter notebook demonstrates how to use the the API (https://github.com/usc-isi-i2/datamart-api/blob/master/Datamart%20Data%20API%20Demo.ipynb).

Datamart repository: https://github.com/usc-isi-i2/datamart-api
T2WML repository: https://github.com/usc-isi-i2/t2wml
Fuzzy search repository: https://github.com/usc-isi-i2/wikidata-fuzzy-search