



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.

The screenshot shows the T2WML version 2.9.4 interface. On the left is a File Tree with folders for 'AMIS.xlsx' and 'fsi-2012.csv'. The main area displays a table with columns A-F and rows 1-41. Row 10 is highlighted, showing 'TOTAL SUPPLY' for 'Argentina' in 'Million tonnes' with a value of '18.10' for the year '2006'. On the right, the 'Output Preview' shows a table with columns A-E and rows 1-16. Row 10 is highlighted, showing 'Argentina' with 'TOTAL SUPPLY' of '18.10' in 'Million tonnes' at '2006-01-01T00:00:00'. Below the preview is a 'YAML Editor' and an 'Output' section showing the structured data for 'Argentina (Q14)'.

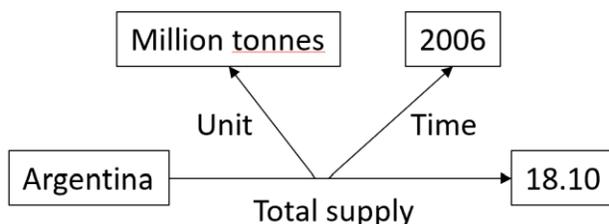
subject	property	value	unit	point in time
Argentina	TOTAL SUPPLY	18.10	Million tonnes	2006-01-01T00:00:00
Argentina	TOTAL SUPPLY	19.94	Million tonnes	2007-01-01T00:00:00
Argentina	TOTAL SUPPLY	12.76	Million tonnes	2008-01-01T00:00:00
Argentina	TOTAL SUPPLY	10.42	Million tonnes	2009-01-01T00:00:00
Argentina	TOTAL SUPPLY	17.14	Million tonnes	2010-01-01T00:00:00
Argentina	TOTAL SUPPLY	17.99	Million tonnes	2011-01-01T00:00:00
Argentina	TOTAL SUPPLY	9.19	Million tonnes	2012-01-01T00:00:00
Argentina	TOTAL SUPPLY	9.95	Million tonnes	2013-01-01T00:00:00
Argentina	TOTAL SUPPLY	16.48	Million tonnes	2014-01-01T00:00:00
Argentina	TOTAL SUPPLY	17.10	Million tonnes	2015-01-01T00:00:00
Argentina	TOTAL SUPPLY	20.51	Million tonnes	2016-01-01T00:00:00
Argentina	TOTAL SUPPLY	19.80	Million tonnes	2017-01-01T00:00:00
Argentina	TOTAL SUPPLY	21.01	Million tonnes	2018-01-01T00:00:00
Argentina	Opening Stocks	3.55	Million tonnes	2006-01-01T00:00:00
Argentina	Opening Stocks	3.45	Million tonnes	2007-01-01T00:00:00

```

Output
-----
Argentina (Q14)
TOTAL SUPPLY (Pcausx3-total_supply_) 18.10 ( Million tonnes)
- point in time: 2006-01-01T00:00:00
  
```

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 Argentina 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

The screenshot shows the ISI Datamart search interface. At the top, there is a search bar with the query 'food' and a 'Choose Regions' button. Below the search bar, a list of regions is displayed: Ethiopia, Abazinia, Afghanistan, Albania, Algeria... A 'Debug queries' button is also visible.

The search results are presented as a grid of dataset tiles. Each tile contains the following information:

- Dataset Name and ID:** e.g., 'Food, beverages and t... (P110763)' with a red score of 0.046.
- Description:** A brief text description of the dataset's content.
- Time Range:** e.g., '1990 - 2015'.
- Time Precision:** e.g., 'year'.
- Count:** e.g., '26 records'.

On the right side of the interface, a time series chart is displayed for the dataset 'P110026 Food production Index (2004-2006 = 100)'. The chart shows the index value over time from 1960 to 2010. The y-axis ranges from 0 to 1000. The x-axis shows years from 1960 to 2010. The chart features a prominent green line for Afghanistan, which peaks at approximately 1000 around 1980. Other countries are represented by various colored lines, mostly clustered below 200. A legend at the bottom identifies the countries: Afghanistan (green), Albania (purple), Algeria (blue), Angola (orange), Antigua and Barbuda (light blue), Argentina (yellow), and Armenia (red).

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>

