Discovering Related Data Sources in Data Portals

Andreas Wagner, Peter Haase,
Achim Rettinger, Holger Lamm

1st International Workshop on Semantic Statistics

Sydney, Oct 22, 2013

Potential of Open (Statistics) Data

BERLIN OPEN DATA

DESTATIS wissen. nutzen.

UNdata A world of information

WORLD BANK

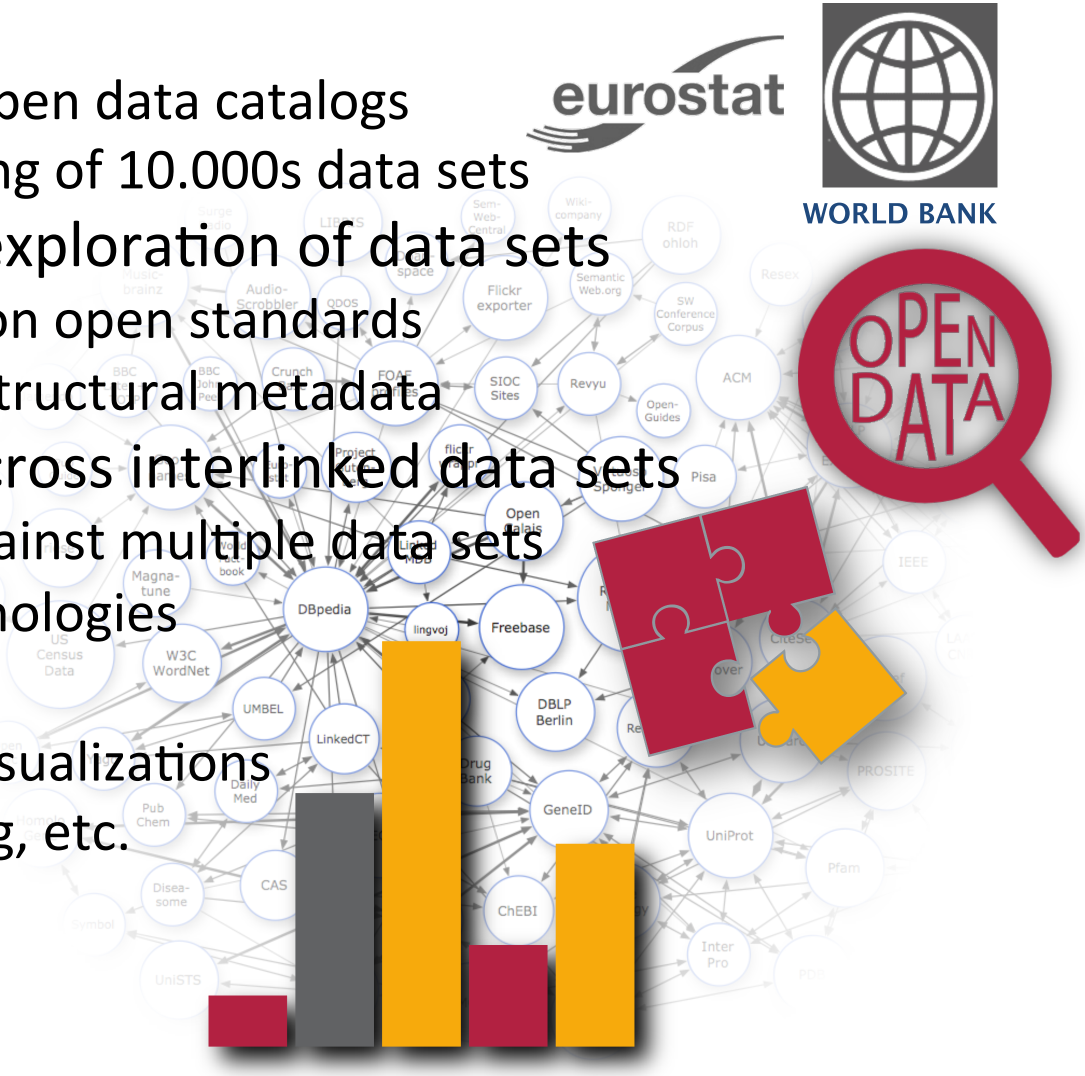
DATA.GOV EMPOWERING PEOPLE

OECD

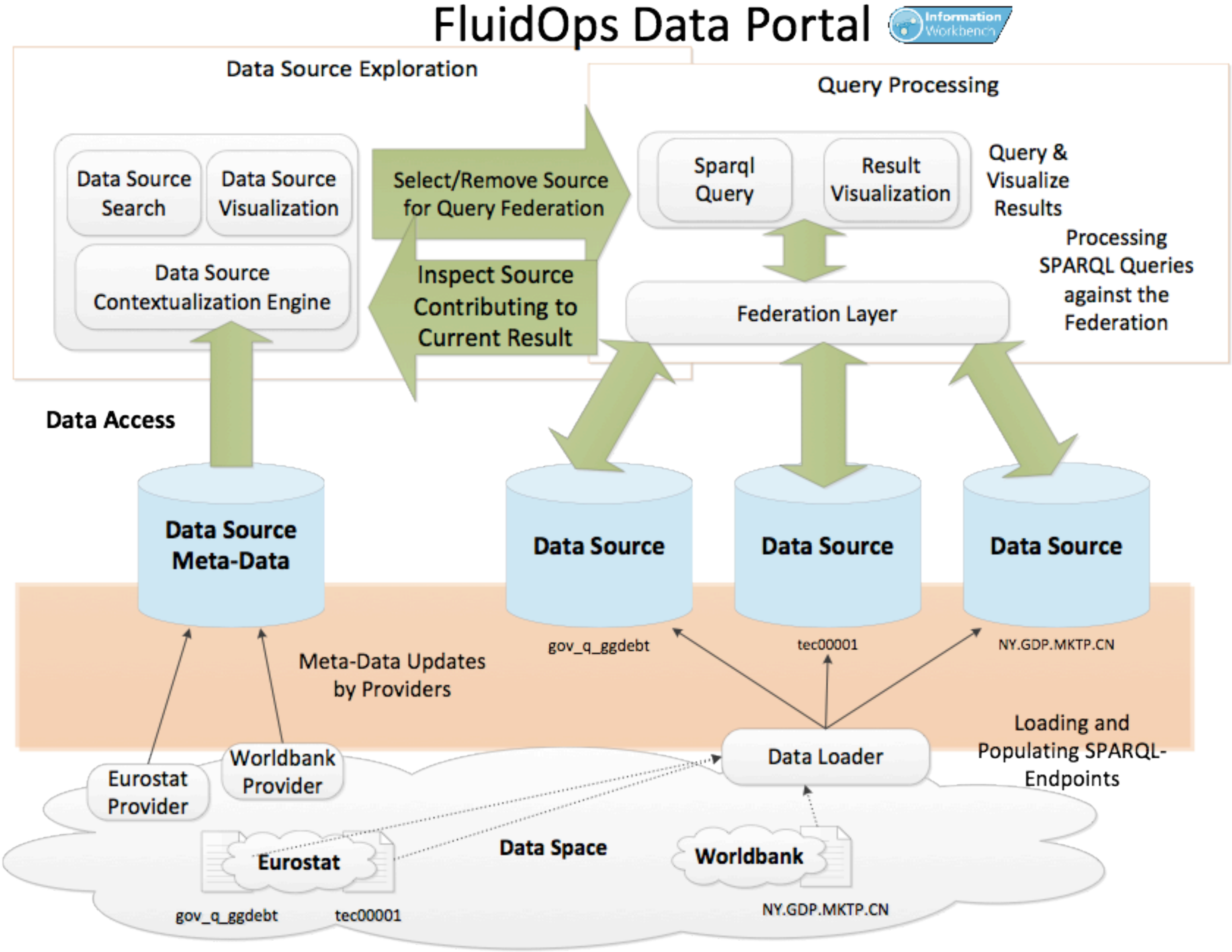
eurostat

fluidOps Open Data Portal

- Data collection
  - Integration of major open data catalogs
  - Automated provisioning of 10.000s data sets
- Portal for search and exploration of data sets
  - Rich metadata based on open standards
  - Both descriptive and structural metadata
- Integrated querying across interlinked data sets
  - Easy to use queries against multiple data sets
  - Using federation technologies
- Self-service UI
  - Custom queries and visualizations
  - Widgets, dashboarding, etc.



FluidOps Data Portal



Data Source Exploration

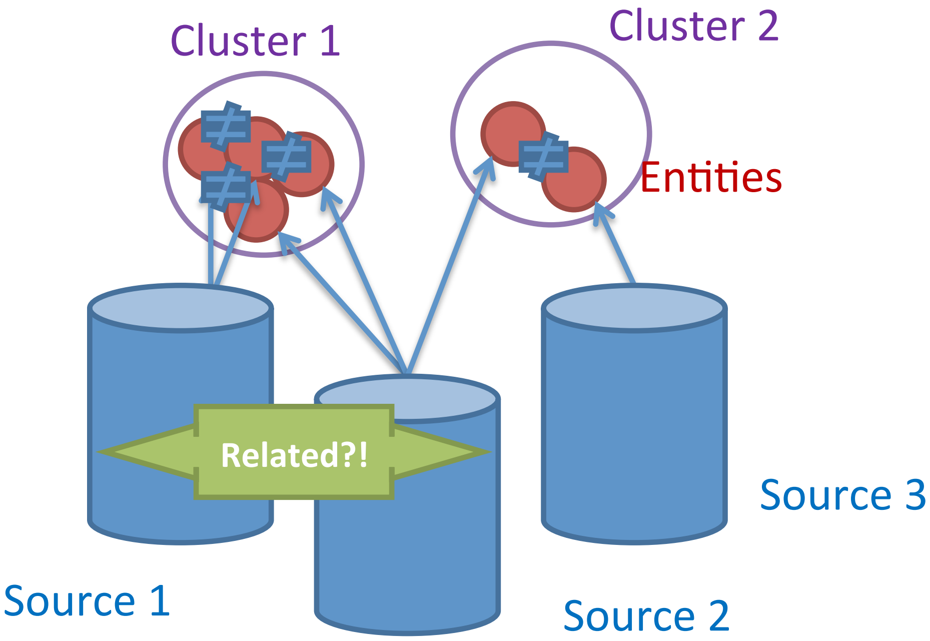
Query Processing

Finding Related Data Sets

- Many information needs require analysis of multiple data sets
- Example: Compare and correlate GDP, population and public debt of countries over time
- Task of finding related data sets
  - Identify data sets that are similar, but complementary
  - To support queries across multiple data sets, e.g. in the form of joins and unions
- Inspiration: Finding related tables
  - Entity complement: same attributes, complementing entities
  - Schema complement: same entities, complementing attributes

Finding Related Data Sources via Related Entities

- Data Model: Data source is a set of multiple RDF graphs
- Intuition: if data sources contain similar entities, they are *somehow* related
- Approach:
  1. Entity Extraction
  2. Entity Similarity
  3. Entity Clustering



Related Entities (2)

1. Entity Extraction
   - Sample over entities in data graphs in D
   - For each entity crawl its surrounding sub-graph [1]
2. Entity Similarity
   - Define dissimilarity measure between two entities based on kernel functions
   - Compare entity structure and literals via different kernels [2,3]
3. Entity Clustering
   - Apply *k*-means clustering to discover similar entities [4]

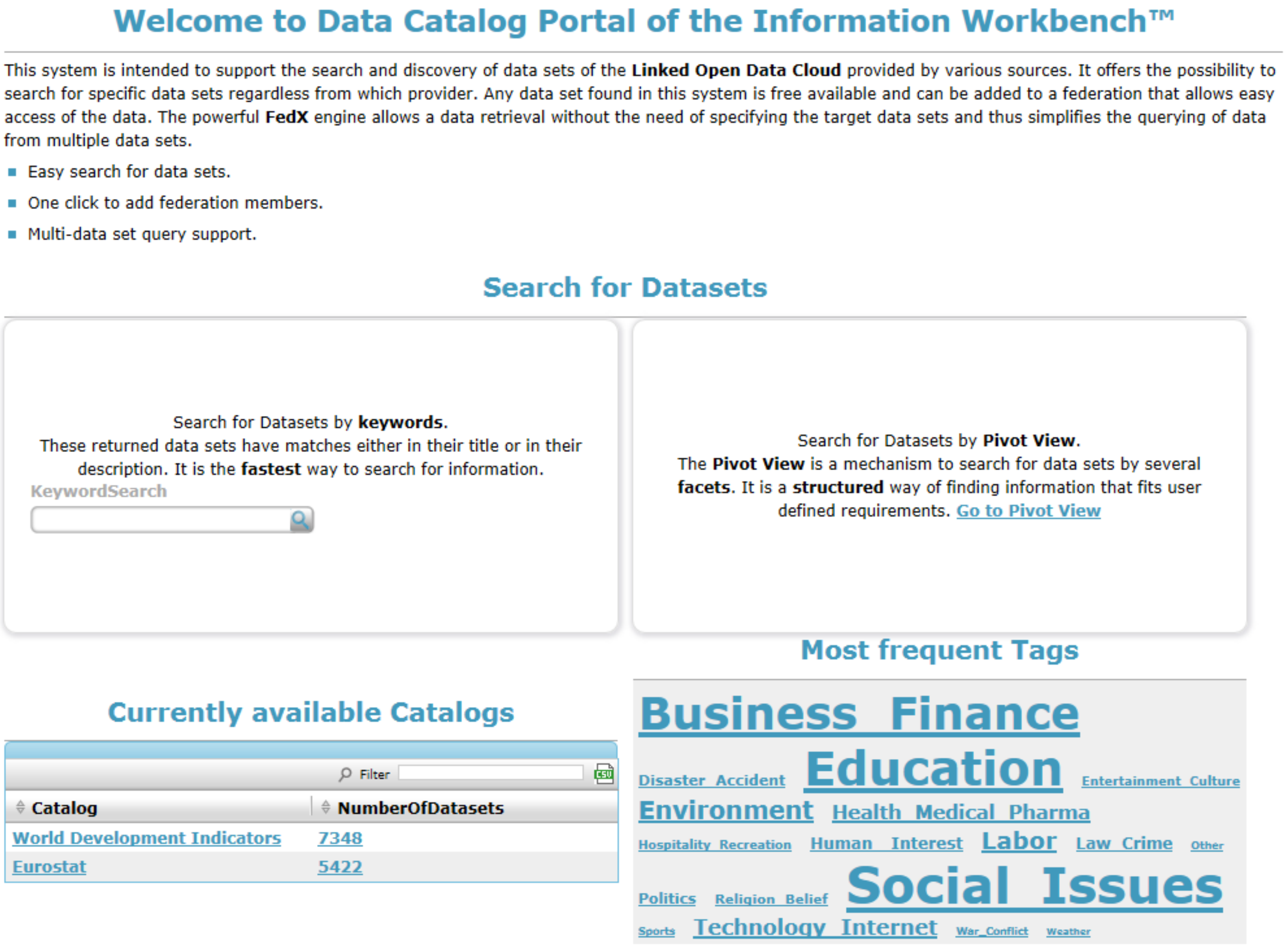
Contextualisation Score

- Contextualization score for data source D’’ given D’: ec(D’’|D’) and sc(D’’|D’)
- *Entity complement* score

$$ec(D'' \mid D') := \sum_{C_j \in \, cluster(D')} \frac{\mathbb{1}(C_j, D'')|C_j|}{|C_j|}$$

- *Schema complement* score

$$sc(D'' \mid D') := \sum_{C_j \in \, cluster(D'')} \frac{|props(C_j) \setminus \bigcup_{C_i \in \, cluster(D')} props(C_i)|}{|props(C_j)|}$$



Welcome to Data Catalog Portal of the Information Workbench™

This system is intended to support the search and discovery of data sets of the **Linked Open Data Cloud** provided by various sources. It offers the possibility to search for specific data sets regardless from which provider. Any data set found in this system is free available and can be added to a federation that allows easy access of the data. The powerful **FedX** engine allows a data retrieval without the need of specifying the target data sets and thus simplifies the querying of data from multiple data sets.

- Easy search for data sets.
- One click to add federation members.
- Multi-data set query support.

Search for Datasets

Search for Datasets by **keywords**.
These returned data sets have matches either in their title or in their description. It is the **fastest** way to search for information.

Search for Datasets by **Pivot View**.
The **Pivot View** is a mechanism to search for data sets by several **facets**. It is a **structured** way of finding information that fits user defined requirements. Go to Pivot View

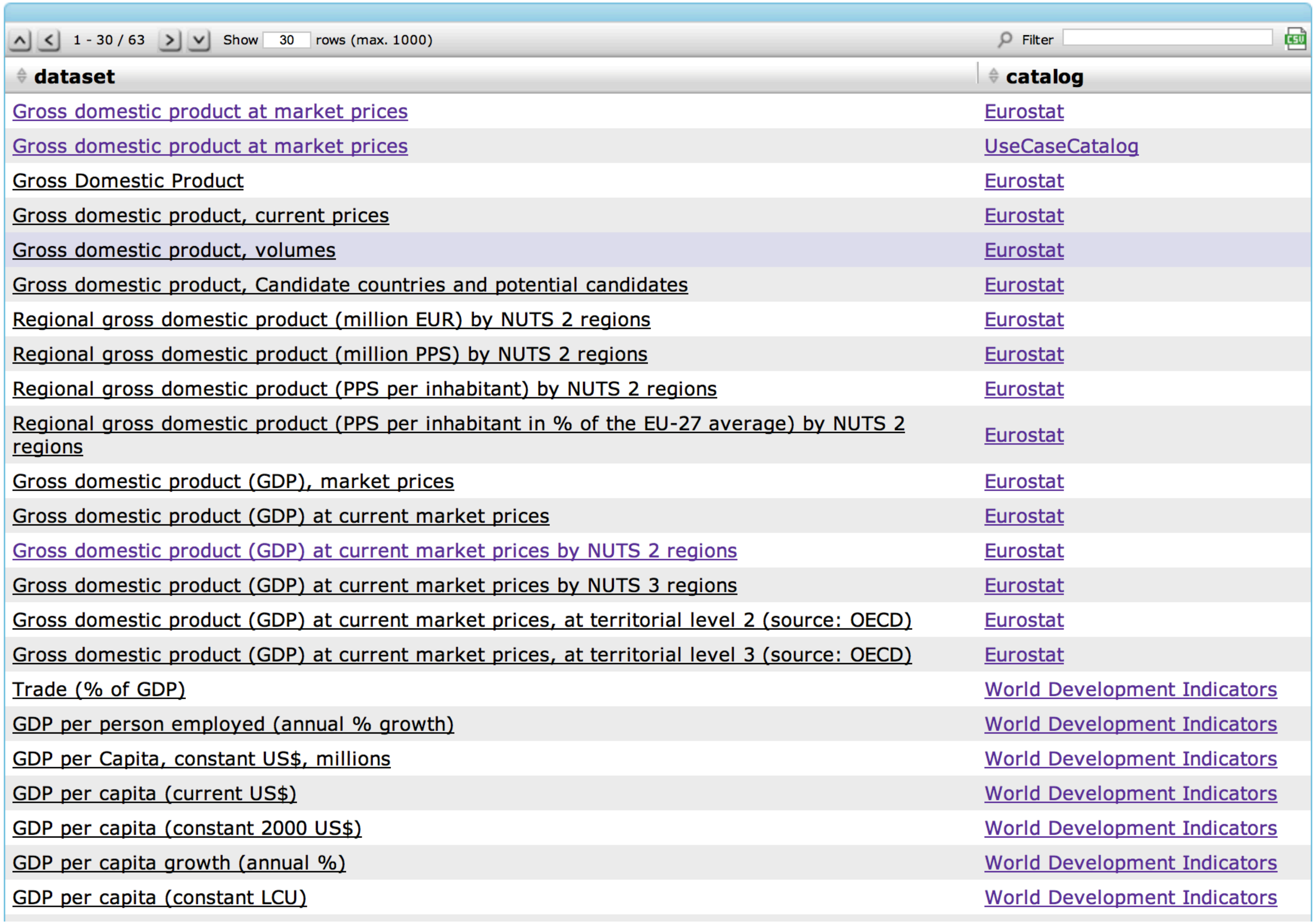
Currently available Catalogs

Catalog	NumberOfDatasets
World Development Indicators	7348
Eurostat	5422

Most frequent Tags

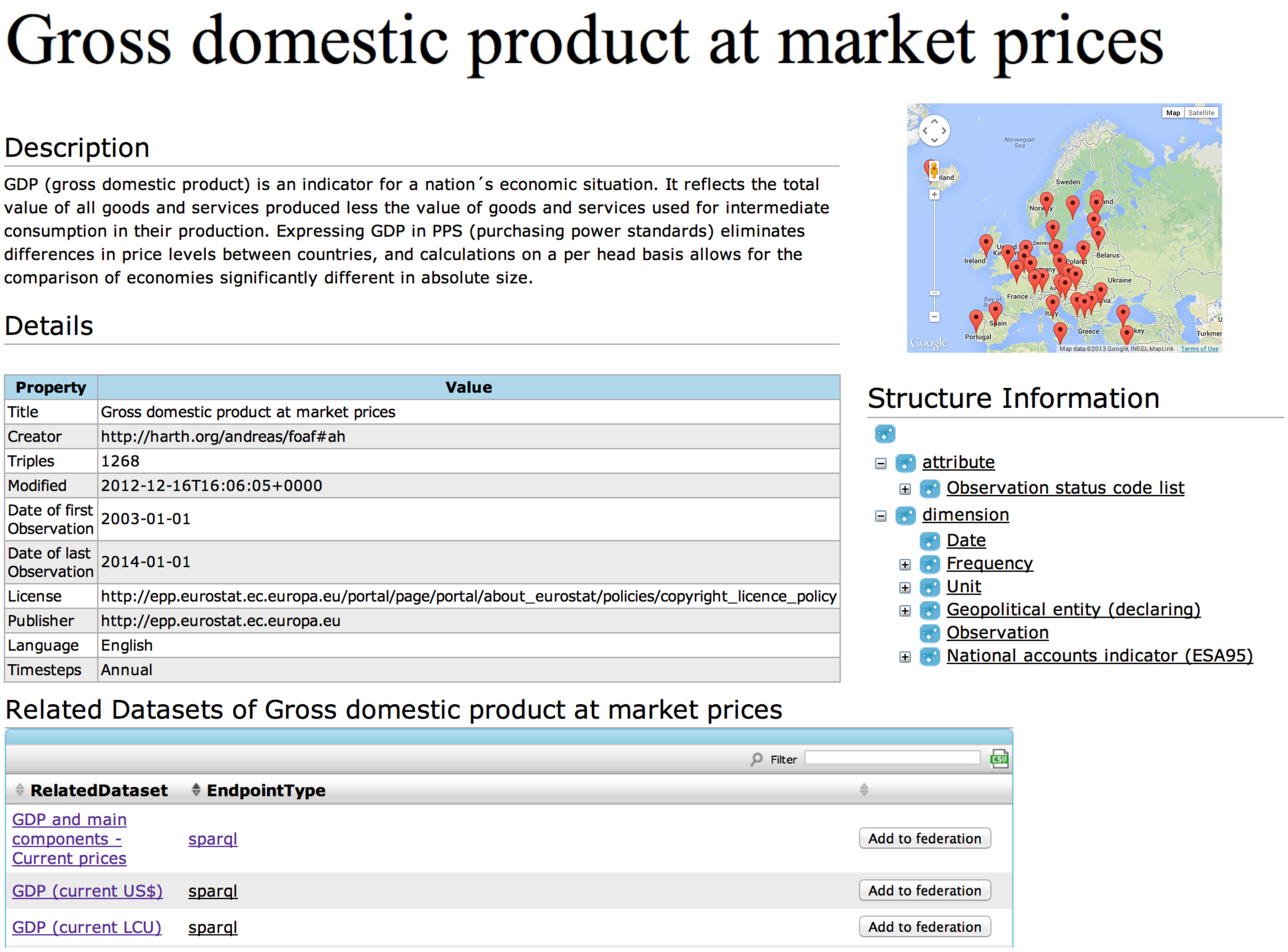
Business_Finance Disaster_Accident Education Entertainment_Culture Environment Health_Medical_Pharma Hospitality_Recreation Human_Interest Labor Law_Crime Other Politics Religion_Belief Social_Issues Sports Technology_Internet War_Conflict Weather

Search for Gross Domestic Product



1 - 30 / 63 Show 30 rows (max. 1000) Filter

dataset	catalog
Gross domestic product at market prices	Eurostat
Gross domestic product at market prices	UseCaseCatalog
Gross Domestic Product	Eurostat
Gross domestic product, current prices	Eurostat
Gross domestic product, volumes	Eurostat
Gross domestic product, Candidate countries and potential candidates	Eurostat
Regional gross domestic product (million EUR) by NUTS 2 regions	Eurostat
Regional gross domestic product (million PPS) by NUTS 2 regions	Eurostat
Regional gross domestic product (PPS per inhabitant) by NUTS 2 regions	Eurostat
Regional gross domestic product (PPS per inhabitant in % of the EU-27 average) by NUTS 2 regions	Eurostat
Gross domestic product (GDP), market prices	Eurostat
Gross domestic product (GDP) at current market prices	Eurostat
Gross domestic product (GDP) at current market prices by NUTS 2 regions	Eurostat
Gross domestic product (GDP) at current market prices by NUTS 3 regions	Eurostat
Gross domestic product (GDP) at current market prices, at territorial level 2 (source: OECD)	Eurostat
Gross domestic product (GDP) at current market prices, at territorial level 3 (source: OECD)	Eurostat
Trade (% of GDP)	World Development Indicators
GDP per person employed (annual % growth)	World Development Indicators
GDP per Capita, constant US$, millions	World Development Indicators
GDP per capita (current US$)	World Development Indicators
GDP per capita (constant 2000 US$)	World Development Indicators
GDP per capita growth (annual %)	World Development Indicators
GDP per capita (constant LCU)	World Development Indicators



Gross domestic product at market prices

Description

GDP (gross domestic product) is an indicator for a nation´s economic situation. It reflects the total value of all goods and services produced less the value of goods and services used for intermediate consumption in their production. Expressing GDP in PPS (purchasing power standards) eliminates differences in price levels between countries, and calculations on a per head basis allows for the comparison of economies significantly different in absolute size.

Details

Property	Value
Title	Gross domestic product at market prices
Creator	http://harth.org/andreas/foaf#ah
Triples	1268
Modified	2012-12-16T16:06:05+0000
Date of first Observation	2003-01-01
Date of last Observation	2014-01-01
License	http://epp.eurostat.ec.europa.eu/portal/page/portal/about_eurostat/policies/copyright_licence_policy
Publisher	http://epp.eurostat.ec.europa.eu
Language	English
Timesteps	Annual

Structure Information

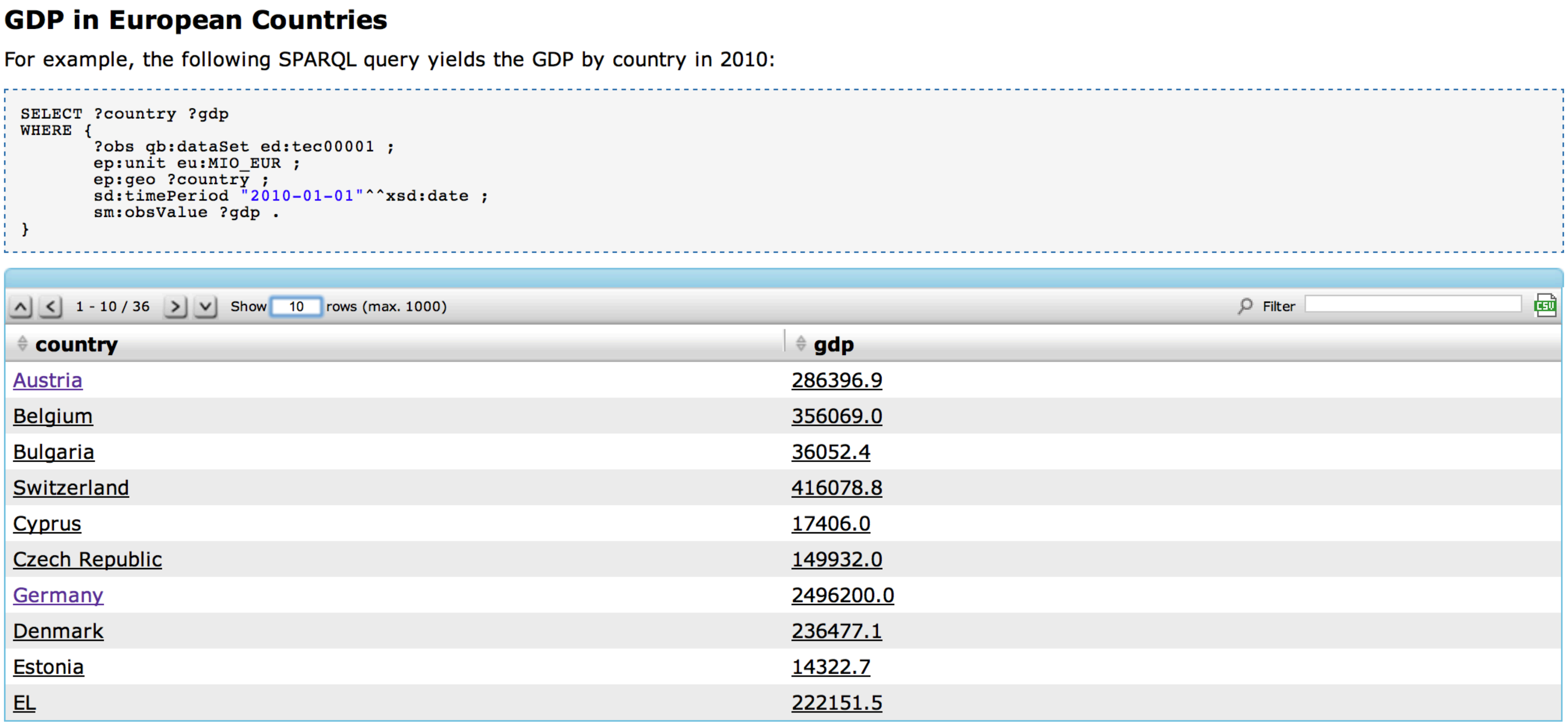
- ⊟ attribute
  - ⊞ Observation status code list
- ⊟ dimension
  - Date
  - ⊞ Frequency
  - ⊞ Unit
  - ⊞ Geopolitical entity (declaring)
  - Observation
  - ⊞ National accounts indicator (ESA95)

Related Datasets of Gross domestic product at market prices

Filter

RelatedDataset	EndpointType	
GDP and main components - Current prices	sparql	Add to federation
GDP (current US$)	sparql	Add to federation
GDP (current LCU)	sparql	Add to federation

Querying the Data Set



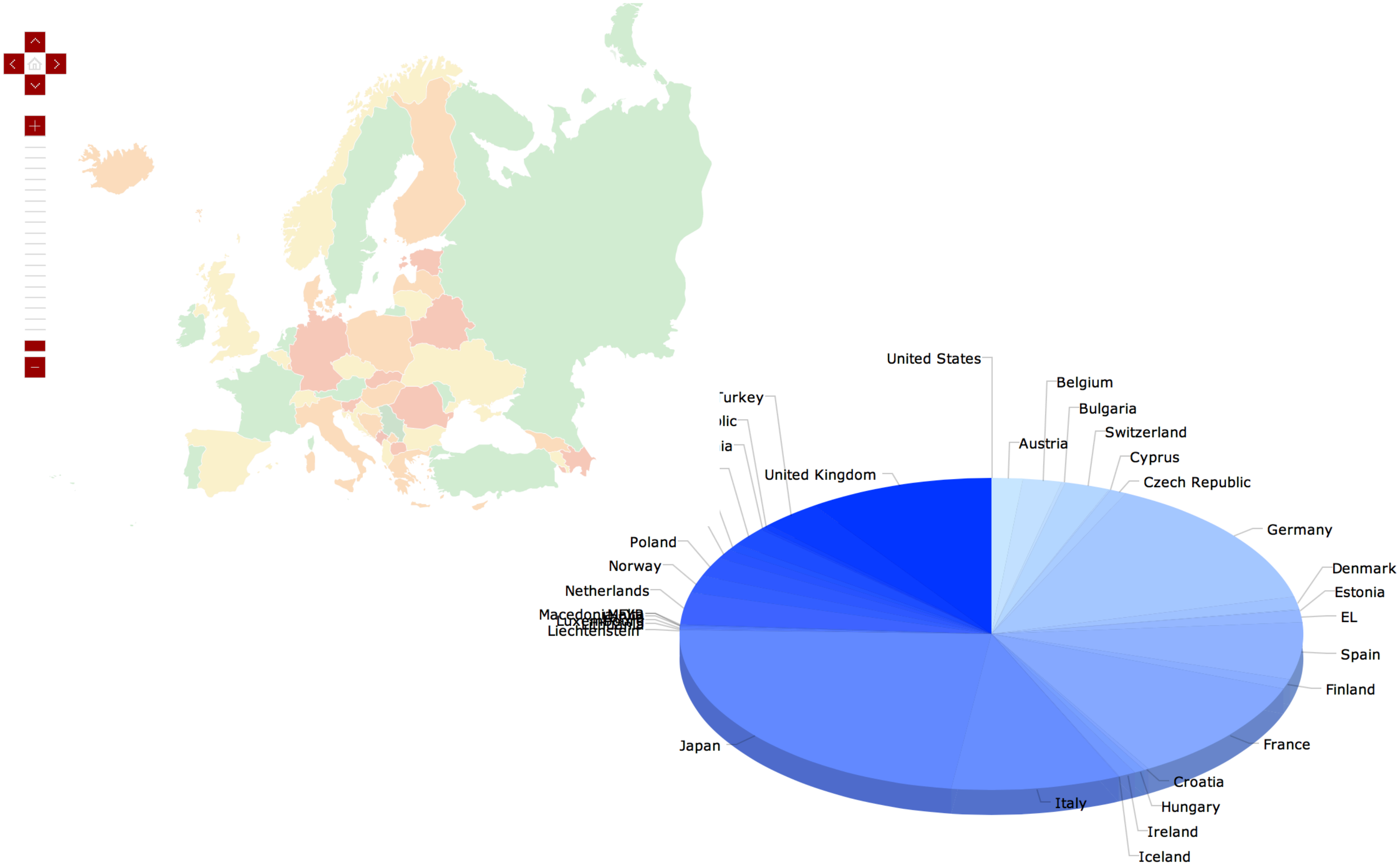
GDP in European Countries

For example, the following SPARQL query yields the GDP by country in 2010:

```
SELECT ?country ?gdp
WHERE {
       ?obs qb:dataSet ed:tec00001 ;
       ep:unit eu:MIO_EUR ;
       ep:geo ?country ;
       sd:timePeriod "2010-01-01"^^xsd:date ;
       sm:obsValue ?gdp .
}
```

country	gdp
Austria	286396.9
Belgium	356069.0
Bulgaria	36052.4
Switzerland	416078.8
Cyprus	17406.0
Czech Republic	149932.0
Germany	2496200.0
Denmark	236477.1
Estonia	14322.7
EL	222151.5

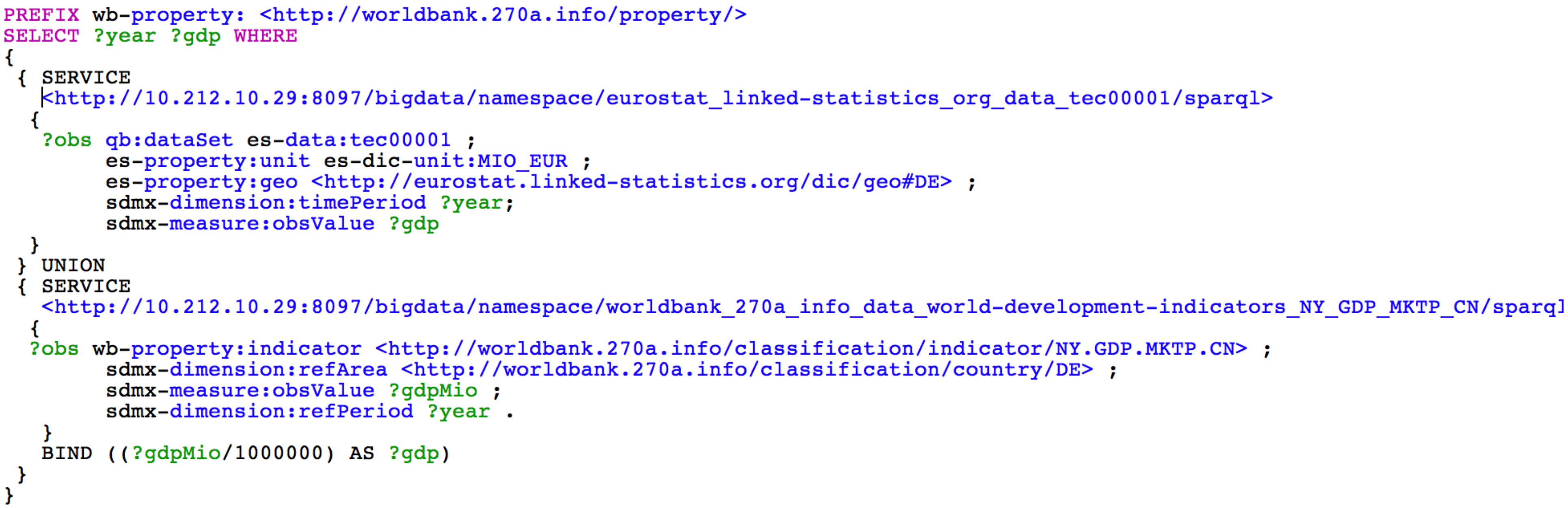
Visualizing the Results



Queries Across Related Data Sets

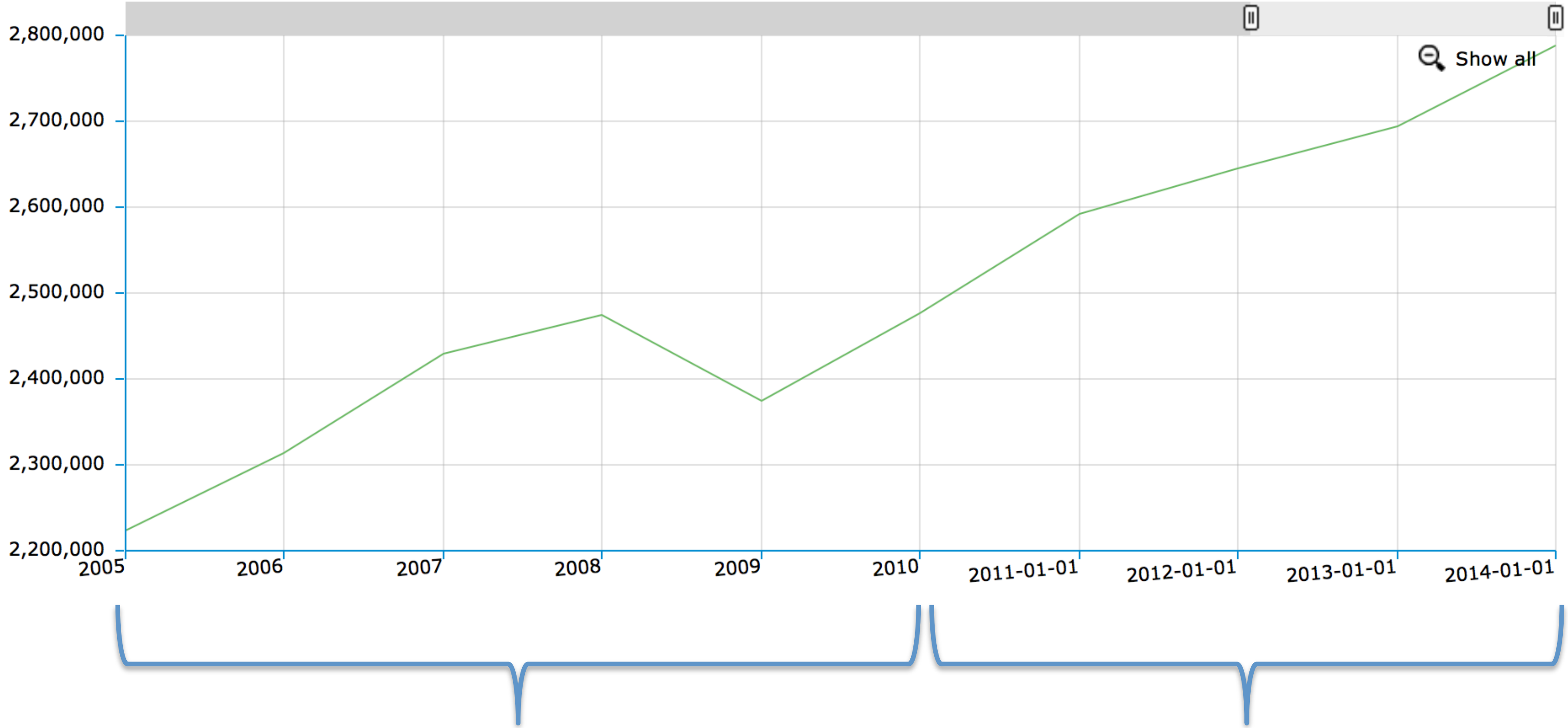
- Query for GDP of Germany

- Union of results from
  - Worldbank: GDP (current US$) (up to 2010)
  - Eurostat: GDP at Market Prices (including projected values until 2014)



```
PREFIX wb-property: <http://worldbank.270a.info/property/>
SELECT ?year ?gdp WHERE
{
 { SERVICE
   <http://10.212.10.29:8097/bigdata/namespace/eurostat_linked-statistics_org_data_tec00001/sparql>
  {
   ?obs qb:dataSet es-data:tec00001 ;
        es-property:unit es-dic-unit:MIO_EUR ;
        es-property:geo <http://eurostat.linked-statistics.org/dic/geo#DE> ;
        sdmx-dimension:timePeriod ?year;
        sdmx-measure:obsValue ?gdp
  }
 } UNION
 { SERVICE
   <http://10.212.10.29:8097/bigdata/namespace/worldbank_270a_info_data_world-development-indicators_NY_GDP_MKTP_CN/sparql
  {
  ?obs wb-property:indicator <http://worldbank.270a.info/classification/indicator/NY.GDP.MKTP.CN> ;
        sdmx-dimension:refArea <http://worldbank.270a.info/classification/country/DE> ;
        sdmx-measure:obsValue ?gdpMio ;
        sdmx-dimension:refPeriod ?year .
   }
   BIND ((?gdpMio/1000000) AS ?gdp)
 }
}
```

Queries Across Related Data Sets



Data from Worldbank

Data from Eurostat

Summary and Outlook

- Techniques for finding related data sets
  - Based on finding related entities
- Implementation available in open data portal

- Outlook
  - Finding relevant related data sources for a given information need
  - End user interfaces for formulating queries across data sets (see Optique project)
  - Operators for combining data cubes
  - Interactive visualization and exploration of combined data cubes (see OpenCube project)

References

[1] G. A. Grimnes, P. Edwards, and A. Preece. **Instance based clustering of semantic web resources.** In ESWC, 2008.

[2] U. Lösch, S. Bloehdorn, and A. Rettinger. **Graph kernels for RDF data**. In ESWC, 2012.

[3] J. Shawe-Taylor and N. Cristianini. **Kernel Methods for Pattern Analysis**. 2004.

[4] R. Zhang and A. Rudnicky. **A large scale clustering scheme for kernel k-means.** In Pattern Recognition, 2002.