LOD2
Creating Knowledge out of Interlinked Data

Publishing Statistical Data As Linked Open Data

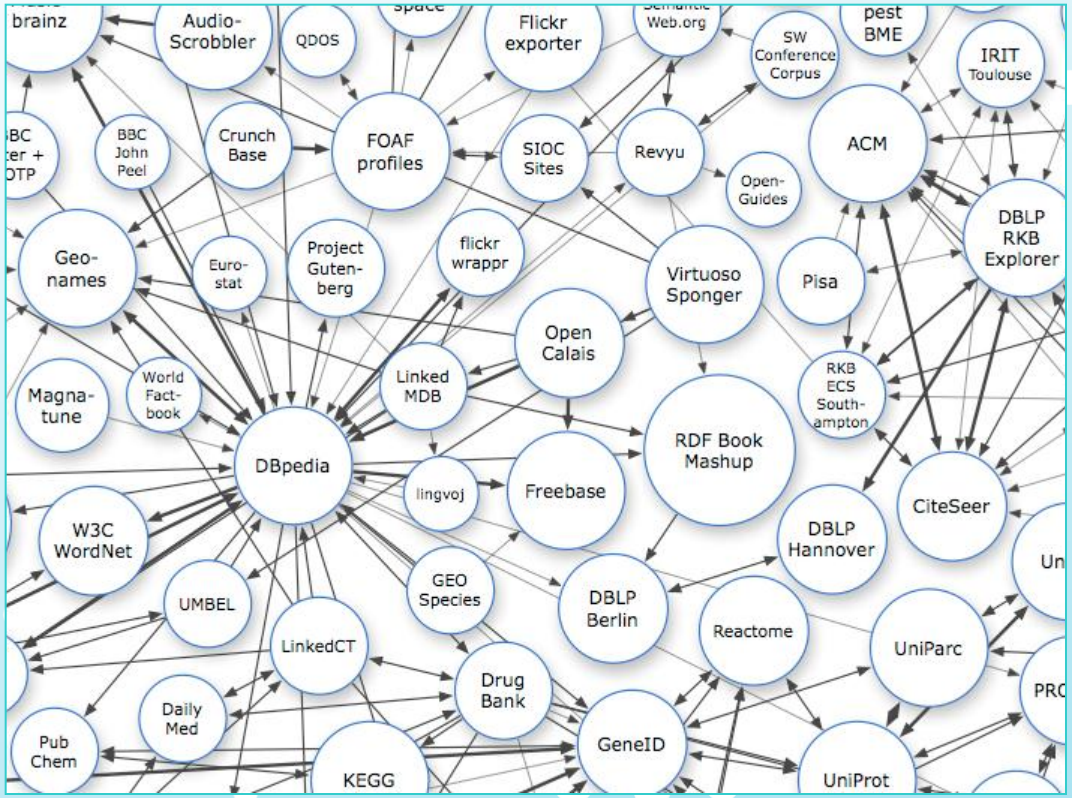
Uroš Milošević, Valentina Janev, Mirko Spasić[1], Jelena Milojković[2], Sanja Vraneš[1]
[1]University of Belgrade, Mihajlo Pupin Institute
[2]Statistical Office of the Republic of Serbia

SEVENTH FRAMEWORK PROGRAMME
institut MIHAJLO PUPIN

http://lod2.eu

A Web of Data

- Huge amounts of data need to be made available in a standard format, reachable and manageable by Semantic Web tools



Statistical Data

- **Statistical offices**
  - Continuous sources of fresh, structured data
  - Lack the means for exposing, sharing, and interlinking this data on the Semantic Web
- Statistical data underpins many of the mash-ups and visualizations we see on the Web
- Statistical data provides the foundations for policy prediction, planning and adjustments

Statistical Data As LOD

- Third party annotations and linking
- Flexible combination of data across both statistical and non-statistical datasets
- New ways of manipulating the data
- Machine readable means of publication that supports an out-of-the-box web API for programmatic access.

LOD2 (Stack)

- A European initiative to
  - improve coherence and quality of data published on the Web
  - close the performance gap between relational and RDF data management
  - establish trust on the Linked Data Web
  - generally lower the entrance barrier for data publishers and users

- **The LOD2 Stack**
  - A distribution of integrated software tools and components enabling corporations, organizations and individuals to employ Linked Data technologies with minimal initial investments

Data Source

- Statistical Office of the Republic of Serbia (SORS)
  - Information published on monthly, quarterly and yearly basis
  - Mostly available as open, downloadable, free of charge documents in PDF format
  - Raw data with short and long-term indicators organized in a central statistics publication database
  - Strong interest in being able to publish statistical data in a web-friendly format to enable it to be linked and combined with related information
- Input data provided as XML files

Statistical Data in RDF

- Linked Data enables datasets to be linked together through references to common concepts
- HTTP URIs (usually) used to name the entities and concepts, so that data consumers are provided with more information, including links to other related URIs
- A mechanism for data publishing on the web which supports easy discovery and cross-linking of published data.

Design Choices: SDMX

- The SORS is in the process of harmonization of standards, classifications and methodologies with the system of official statistics of the EU, and, therefore, coordinates its activities in accordance with the European Statistics Code of Practice

- **The Statistical Data and Metadata eXchange (SDMX)**
  - An ISO standard used by U.S. Federal Reserve Board, the European Central Bank, Eurostat, the WHO, the IMF, and the World Bank
  - Allows aggregation across national boundaries
  - **Not** web-friendly

Design Choices: Data Cube

- RDF vocabulary focused purely on the publication of multi-dimensional data on the Web
- Supports extension vocabularies to enable publication of other aspects of statistical data flows
- Compatible with the SDMX information model

- **SDMX-RDF**
  - Extension vocabulary that provides a layer on top of Data Cube to describe domain semantics, dataset's metadata, and other crucial information needed in the process of statistical data exchange

Design Choices: Data Cube

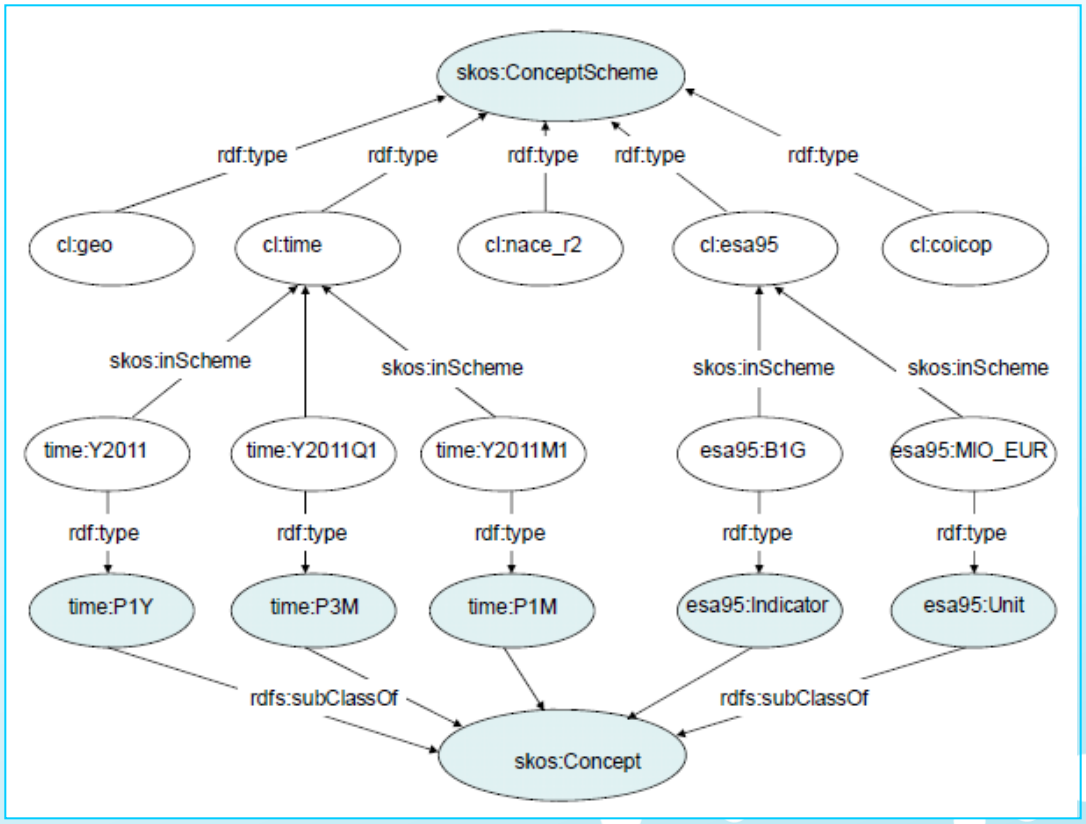
- Dimensions, attributes and measures represented as instances of the abstract `qb:ComponentProperty` class, i.e. one of its sub-classes
  - `qb:DimensionProperty`
  - `qb:AttributeProperty`
  - `qb:MeasureProperty`

- Properties used to describe
  - the concept being represented (e.g. time or geographic area)
  - the nature of the component (dimension, attribute or measure),
  - the type or code list used to represent the value.

Design Choices: Code Lists

- Eurostat requires using standard classification schemas such as NACE, COICOP, PRODCOM etc.
- National statistical offices use also local coding systems, applicable in their countries
- Well defined code lists facilitate dataset comparison, interlinking, discovery and merging

Design Choices: Code Lists



- Concepts represented as `skos:Concept`
- Concepts grouped in concept schemes that serve as code lists (`skos:ConceptScheme`) from which the dataset dimensions draw on their values

Transforming the Data

- Statistical data in XML form is passed as input to the XSLT processor and transformed into RDF using the appropriate vocabularies and concept schemes

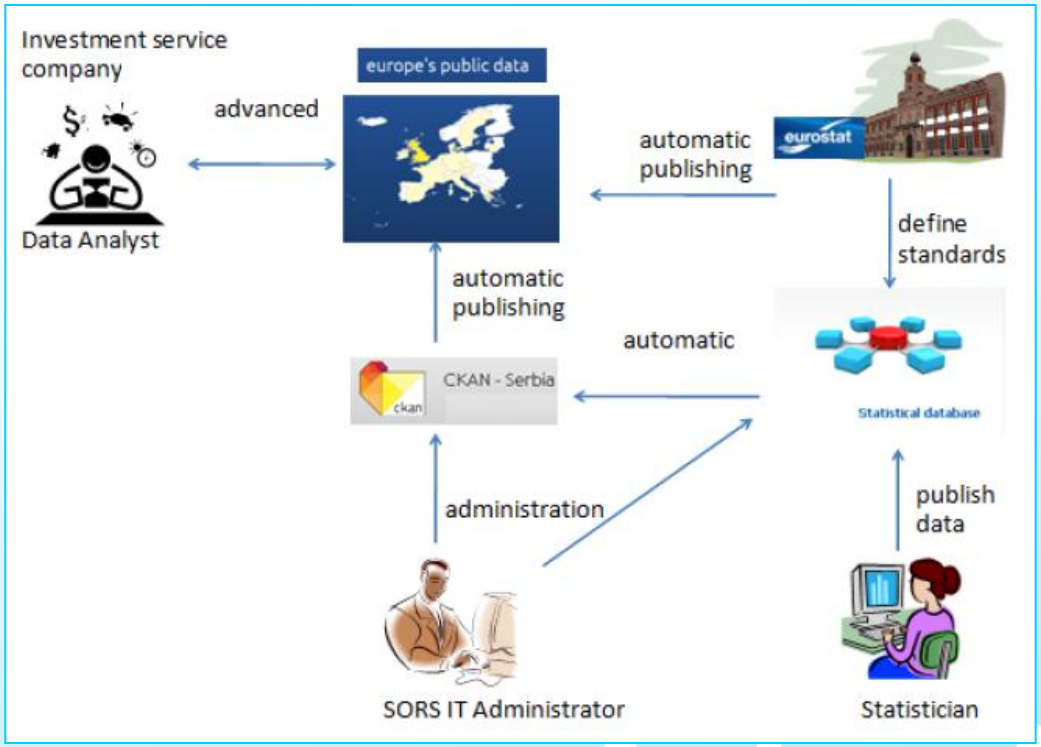
```
<rs:geo>
  <xsl:attribute name="rdf:resource">
  <xsl:variable name="region" select="current()/attribute::Teritory"/>
  <xsl:variable name="geo" select="$geo_cl//
                rs-geo:Region[rdfs:label=$region]/attribute::rdf:ID" />
  <xsl:value-of select="concat('RS-DIC/geo#',$geo)" />
</xsl:attribute>
</rs:geo>
```

Transforming the Data

```
<qb:Observation rdf:about="RS-DATA/NA/GVA/data#obs978">
   <qb:dataSet rdf:resource="RS-DATA/NA/GVA/data"/>
   <sdmx-attribute:unitMeasure rdf:resource="RS-DIC/esa95-unit#MIO_NAT"/>
   <rs:geo rdf:resource="RS-DIC/geo#RS"/>
   <rs:time rdf:resource="RS-DIC/time#Y2001Q3"/>
   <sdmx-measure:obsValue>
      183802.6
   </sdmx-measure:obsValue>
</qb:Observation>
```

A sample transformation result for a single observation in RDF/XML syntax

SORS Workflow Example



- **Serbian CKAN open government metadata repository set up as part of the PublicData.eu data hub network.**

Results

- Guidelines for helping national statistics offices in the process of moving from (local) raw statistical data to (globally visible) rich collections of interrelated statistical datasets
- Data represented using a future-proof novel method, compatible with international statistical standards
- Resulting data relies on both international and domestic code lists
- Results cataloged in a local metadata repository
- Periodical “harvesting” at an international level scheduled, thereby increasing transparency and improving public service delivery, while enriching the Linked Data cloud

LOD2

Creating Knowledge out of Interlinked Data

Questions?