Lifting Open Data Portals to the Data Web

Sander van der Waal[1(✉)], Krzysztof Węcel[2(✉)], Ivan Ermilov[3], Valentina Janev[4], Uroš Milošević[4], and Mark Wainwright[1]

[1] Open Knowledge Foundation, Cambridge, UK
{sander.vanderwaal,mark.wainwright}@okfn.org
[2] I2G, Poznań, Poland
krzysztof.wecel@i2g.pl
[3] University of Leipzig, Leipzig, Germany
iermilov@informatik.uni-leipzig.de
[4] Institute Mihajlo Pupin, Belgrade, Serbia
{valentina.janev,uros.milosevic}@pupin.rs

Abstract. Recently, a large number of open data repositories, catalogs and portals have been emerging in the scientific and government realms. In this chapter, we characterise this newly emerging class of information systems. We describe the key functionality of open data portals, present a conceptual model and showcase the pan-European data portal PublicData.eu as a prominent example. Using examples from Serbia and Poland, we present an approach for lifting the often semantically shallow datasets registered at such data portals to Linked Data in order to make data portals the backbone of a distributed global data warehouse for our information society on the Web.

1 Public Data and Data Portals

Although there are many different sources of data, government data is particularly important because of its scale, breadth, and status as the canonical source of information on a wide range of subjects. Governments gather data in many areas: demographics, elections, government budgets and spending, various types of geospatial data, environmental data, transport and planning, etc. While the data is gathered to support the functions of government, it is increasingly recognised that by publishing government data under permissive open licences (with due precautions to avoid publishing sensitive or personal data), huge amounts of value can be unlocked.

The growth of open government data has been particularly striking in Europe. The EU recognised its advantages very early, and issued the PSI (Public Sector Information) Directive in December 2003. This encouraged governments to make their data available, without restrictions on its use. However, though forward-looking at the time, the directive did allow for charging for use of data, provided that the charges did not exceed those calculated on a cost-recovery basis. It therefore did not require what would now be considered 'open' data. The Directive was revised in 2013, bringing more public bodies within scope and encouraging

S. Auer et al. (Eds.): Linked Open Data, LNCS 8661, pp. 175–195, 2014.
DOI: 10.1007/978-3-319-09846-3_9

free or marginal-cost, rather than recovery-cost, pricing – reflecting what was by then already practice in many EU states. A study in 2011 for the EU estimated the economic value of releasing public sector data throughout the EU at between 30–140 billion EUR.

Except the economical importance, there are additional issues concerning public data that we briefly characterise below: discoverability, harvesting, interoperability, and community engagement.

Discoverability. One of the first problems to be solved when working with any data is where to find it. In using data, one needs exactly the right dataset – with the right variables, for the right year, the right area, etc. – and web search engines, while excellent at finding documents relevant to a given term, do not have enough metadata to find datasets like this, particularly since their main use case is for finding web pages rather than data. There is little point in publishing data if no-one can find it, so how are governments to make their data 'discoverable'? One possibility would be to link it from an on-line tree-structured directory, but such structures are hard to design and maintain, difficult to extend, and do not really solve the problem of making nodes findable when there is a very large number of them (governments typically have at least tens of thousands of datasets).

To solve this problem of discoverability, in the last few years, an increasing number of governments have set up data portals, specialised sites where a publishing interface allows datasets to be uploaded and equipped with high-quality metadata. Using this metadata, users can then quickly find the data they need with searching and filtering features. One good example is the European Open Data portal[1], which is developed by LOD2 partners, using LOD2 stack tools. Numerous countries, including a good number of EU member states, have followed, along with some local (e.g. city) governments.

Harvesting. Many of these portals use CKAN[2], a free, open-source data portal platform developed and maintained by Open Knowledge. As a result they have a standard powerful API, which raises the possibility of combining their catalogues to create a single Europe-wide entry point for finding and using public data. This has been done as part of the LOD2 project: the result is PublicData.eu[3], a data portal also powered by CKAN which uses the platform's 'harvesting' mechanism to copy metadata records of many thousands of datasets from government data portals in over a dozen countries, with new ones added when they become available. Some other portals are also harvested (e.g. city level portals or community-run catalogues of available government data). Sites are regularly polled for changes, ensuring that the aggregate catalogue at PublicData.eu stays roughly in sync with the original catalogues. The PublicData.eu portal is described in more detail in Sect. 2.

[1] http://open-data.europa.eu
[2] http://ckan.org
[3] http://publicdata.eu

Interoperability. Non-CKAN portals can also be harvested if they provide a sufficiently powerful API, but for each different platform, some custom code must be written to link the platform's API to CKAN's harvesting mechanism. A few such sites are included in those harvested by PublicData.eu, but rather than writing endless pieces of code for custom harvesting, effort has instead been directed to working towards defining a standard simple interface which different data catalogues can use to expose their metadata for harvesting. This work in progress can be seen at http://spec.datacatalogs.org.

Community Engagement. If governments want to realise the potential benefits of open data, it is not enough just to publish data and make it discoverable. Even the most discoverable data will not be actually discovered if no-one knows that it exists. It is therefore recognised that best practice in data publishing includes an element of 'community engagement': not just waiting for potential users to find data, but identifying possible re-users, awareness raising, and encouraging re-use.

2 Using PublicData.eu

2.1 Data Publishing

Since PublicData.eu harvests datasets from national portals, it is not used directly by civil servants to publish data. Rather, an entire portal, such as the Romanian portal data.gov.ro, is added to PublicData.eu as a 'harvest source', after which it will automatically be regularly polled for updates. If the portal being added is CKAN-based, then the process of setting up harvesting for it takes only a few minutes. However, it is worth briefly explaining the process of publishing data on a source portals. Government employees have individual accounts with authorisations depending on their department. A web user interface guides an authorised user through the process of publishing data, either uploading or linking to the data itself and adding metadata such as title, description, keywords, licence, etc., which enable users to find and identify it. A published dataset may have a number of data files and displays the publishing department as well as the other added metadata. For most common data formats (including CSV, XLS, PDF, GeoJSON and others), users can not only see the metadata but also preview the data in tables, maps and charts before deciding to download it. The process described is for using a CKAN portal, but other data portals will have similar functionality.

Importantly, publishers are not constrained to publish only linked data, or only data in a particular format. While publication of linked data may be desirable, it is far more desirable for data to be published than to be not published. Pragmatically, this often means governments making data available in whatever form they have. One of the ways in which PublicData.eu adds value to published data is with a feature (described in more detail in Sect. 3) to lift data in the simple CSV spreadsheet format into linked data's native RDF. Two projects by governments to publish well-modelled, high-quality linked data are also described later in this chapter.

2.2 Data Consumption

Users on PublicData.eu can search for data from any of the harvested portals[4]. A search is made by specifying particular words or terms. CKAN searches for matching terms anywhere in a metadata record, including for example the description, and returns results (datasets) ordered by relevance or date modified. The user can add filters to the search to find only datasets from particular countries, or with particular tags or available file formats. A top-level set of nine topic areas (Finance & Budgeting, Transport, Environment, etc.) is also displayed and can be used for filtering search results. As with the source portals described above, they can preview the data in various ways before deciding to download it. Users can search, preview and download without registering, but registering enables a user to follow particular datasets or topics, being notified when the dataset changes or when new datasets are published.

The user interface of CKAN has been translated into a wide range of languages, and users can choose the language in which they interact with the site.

Like all CKAN instances, PublicData.eu can be accessed via an RPC-style API[5], as well as via the web interface. Metadata for each dataset is also available as linked data, in N3 or RDF-XML format[6]. It is also possible to get a dump of the entire catalogue as linked data. As mentioned above, the site also includes a feature for lifting data published in CSV spreadsheets to RDF.

3 Semantic Lifting of CSV to RDF

3.1 Lifting the Tabular Data

Integrating and analysing large amounts of data plays an increasingly important role in today's society. Often, however, new discoveries and insights can only be attained by integrating information from dispersed sources, which requires considerable amounts of time and can be error prone if information is stored in heterogeneous representations.

The Semantic Web and Linked Data communities are advocating the use of RDF and Linked Data as a standardized data publication format facilitating data integration and visualization. Despite its unquestionable advantages, only a tiny fraction of open data is currently available as RDF. At the Pan-European data portal PublicData.eu, which aggregates dataset descriptions from numerous other European data portals, only 1,790 out of more than 49,000 datasets (i.e. just 4 %) were available as RDF. This can be mostly attributed to the fact, that publishing data as RDF requires additional effort in particular with regard to identifier creation, vocabulary design, reuse and mapping.

[4] As of June 2014, there was 1 pan-European, 13 national level, 15 regional data portals, where some of the national portals where community-supported.

[5] http://docs.ckan.org/en/latest/api/index.html

[6] For example http://publicdata.eu/dataset/chemical-water-quality-by-distance-1990to2006.rdf.

Various tools and projects (e.g. Any23, Triplify, Tabels, Open Refine) have been launched aiming at facilitating the lifting of tabular data to reach semantically structured and interlinked data. However, none of these tools supported a truly incremental, pay-as-you-go data publication and mapping strategy, which enabled effort sharing between data owners and consumers. The lack of such an architecture of participation with regard to the mapping and transformation of tabular data to semantically richer representations hampers the creation of an ecosystem for open data publishing and reuse. In order to realize such an ecosystem, we have to enable a large number of potential stakeholders to effectively and efficiently collaborate in the data lifting process. Small contributions (such as fine-tuning of a mapping configuration or the mapping of an individual column) should be possible and render an instant benefit for the respective stakeholder. The sum of many such small contributions should result in a comprehensive Open Knowledge space, where datasets are increasingly semantically structured and interlinked.

The approach presented in this section supports a truly incremental, pay-as-you-go data publication, mapping and visualization strategy, which enables effort sharing between data owners, community experts and consumers. The transformation mappings are crowd-sourced using a Semantic MediaWiki[7] and thus allow incremental quality improvement. The transformation process links related tabular data together and thus enables the navigation between heterogeneous sources. For visualization, we integrate CubeViz for statistical data and Facete for spatial data, which provide the users with the ability to perform simple data exploration tasks on the transformed tabular data. The application of our approach to the PublicData.eu portal results in 15,000 transformed datasets amounting to 7.3 Billion triples[8], thus adding a sizeable part to the Web of Data.

3.2 Tabular Data in PublicData.eu

At the time of writing (May 2014) PublicData.eu comprised 20,396 datasets. Each dataset can comprise several data resources and there are overall 60,000+ data resources available at PublicData.eu. These include metadata such as categories, groups, license, geographical coverage and format. Comprehensive statistics gathered from the PublicData.eu are described in [3].

A large part of the datasets at PublicData.eu (approx. 37 %) are in tabular format, such as, for example, CSV, TSV, XLS, XLSX. These formats do not preserve much of the domain semantics and structure. Also, tabular data represented in the above mentioned formats can be syntactically quite heterogeneous and leaves many semantic ambiguities open, which make interpreting, integrating and visualizing the data difficult. In order to support the exploitation of tabular data, it is necessary to transform the data to standardized formats facilitating the semantic description, linking and integration, such as RDF.

[7] http://wiki.publicdata.eu/

[8] The dynamic dump is available at http://datahub.io/dataset/publicdata-eu-rdf-data.

Other formats represented on the PublicData.eu portal comprise: 42 of the datasets have no format specified, 15 % are human-readable representations (i.e. HTML, PDF, TXT, DOC), the other 6 % are geographical data, XML documents, archives as well as various proprietary formats. Thus for a large fraction (i.e. 42 %) of the datasets a manual annotation effort is required, and at the time of writing they can not be converted automatically due to the absence of the format descriptions. Discussion of the conversion of human-readable datasets (i.e. 15 %) to RDF is out of scope of this book. The known fact is that such conversion has been proven to be time-consuming and error-prone. The other 6 % of the datasets are tackled partially in other projects, for instance, GeoKnow project[9] is aimed at converting geographical data to RDF, whereas statistical data from XML documents are converted within Linked SDMX project[10].

3.3 User-Driven Conversion Framework

The completely automatic RDF transformation as well as the detection and correction of tabular data problems is not feasible. In [3] we devised an approach where the effort is shared between machines and human users. Our mapping authoring environment is based on the popular MediaWiki system. The resulting mapping wiki located at wiki.publicdata.eu operates together with PublicData.eu and helps users to map and convert tabular data to RDF in a meaningful way. To leverage the wisdom of the crowd, mappings are created automatically first and can then be revised by human users. Thus, users improve mappings by correcting errors of the automatic conversion and the cumbersome process of creating mappings from scratch can be avoided in most cases. An overview of the entire application is depicted in Fig. 1.

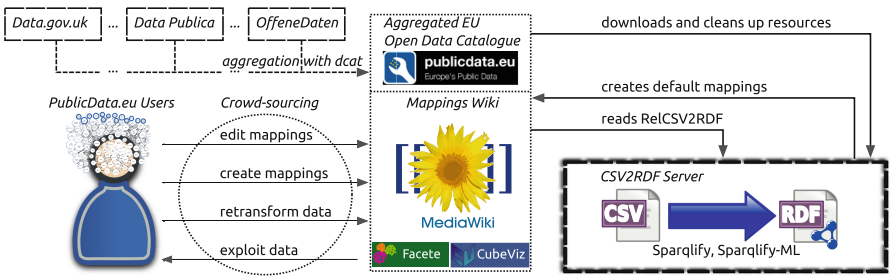

Fig. 1. Architecture of our CSV2RDF extension for PublicData.eu.

Our application continuously crawls CSV resources from PublicData.eu and validates them. Around 20 % of CSV resources are filtered out, mostly because of response timeouts, server errors or missing files. After the validation default

[9] http://geoknow.eu
[10] http://csarven.ca/linked-sdmx-data

mappings are created and resources are converted to RDF. In order to obtain an RDF graph from a table T we essentially use the table as class approach [1], which generates triples as follows: subjects are generated by prefixing each row's ID (in the case of CSV files this by default is the line number) with the corresponding CSV resource URL. The headings become properties in the ontology namespace. The cell values then become the objects. Note that we avoid inferring classes from the CSV file names, as the file names often turned out to be simply labels rather than meaningful type names.

Conversion to RDF is performed by the Sparqlify-CSV. Although the Sparqlify-ML syntax should not pose any problems to users familiar with SPARQL, it is too complicated for novice users and therefore less suitable for crowd-sourcing. To lower the barrier, we define a simplified mapping format, which releases users from dealing with the Sparqlify-ML syntax. Our format is based on MediaWiki templates and thus seamlessly integrates with MediaWiki. To define mappings we created a template called RelCSV2RDF. The complete description for the template is available on the mapping wiki.

At the end of the transformation a page is created for each resource on the mappings wiki at wiki.publicdata.eu. The resource page comprises links to the corresponding resource and dataset on PublicData.eu as well as one or several mappings and visualization links. Each mapping is rendered using the RelCSV2RDF template into a human-readable description of the parameters including links for transformation rerun and RDF download.

Sharing the effort between the human users and machines is never a simple task. The trade-off between human involvement and machine automatic processing should be balanced in a way, that the most precision is achieved with the least time expense from the user side. After automatic mapping generation and resource conversion, user is supposed to find the relevant RDF schema for the given CSV table with third-party tools such as LOV search engine. This task required the background knowledge in the field of Semantic Web, that is the knowledge about existence of specific RDF processing tools. To eliminate this requirement we developed special interface for the finding relevant properties for linking table schema to existing RDF terms.

Additionally, the mapping wiki uses the Semantic MediaWiki [4] (SMW) extension, which enables semantic annotations and embedding of search queries over these annotations within wiki pages. The RelCSV2RDF template utilizes SMW and automatically attaches semantic links (using has_property) from mappings to respective property pages. This allows users to navigate between dataset resources which use the same properties, so that dataset resources are connected through the properties used in their mappings.

3.4 Conversion Results

We downloaded and cleaned 15,551 CSV files, that consume in total 62 GB of disk space. The vast majority (i.e. 85 %) of the published datasets have a size less than 100 kB. A small amount of the resources at PublicData.eu (i.e. 14.5 %) are between 100 kB and 50 MB. Only 44 resources (i.e. 0.5 %) are large and very

large files above 50 MB, with the largest file comprising 3.3 GB. As a result, the largest 41 out of the 9,370 converted RDF resources account for 7.2 (i.e. 98.5 %) out of overall 7.3 billion triples.

The results of the transformation process are summarized in Table 1. Our efficient Sparqlify RDB2RDF transformation engine is capable to process CSV files and generate approx. 4.000 triples per second on a quad core 2.2 GHz machine. As a result, we can process CSV files up to a file size of 50 MB within a minute. This enables us to re-transform the vast majority of CSV files on demand, once a user revised a mapping. For files larger than 50 MB, the transformation is currently queued and processed in batch mode.

Table 1. Transformation results summary

CSV res. converted	9,370	Avg. no. properties per entity	47
CSV res. volume	33 GB	Generated default mappings	9,370
No. generated triples	7.3 billions	Overall properties	80,676
No. entity descriptions	154 millions	Distinct properties	13,490

4 Statistical Data in Serbia

The National Statistical Office is a special professional organisation in the system of state administration that performs expert tasks related to: organisation and conduction of statistical surveys, preparing and adopting unique statistical standards; cooperation with international organisations in order to provide standardisation and data comparability (e.g. EUROSTAT[11]), establishment and maintenance of the system of national accounts, cooperation and expert coordination with bodies and organisations that are in charge of carrying out the statistical surveys, as well as other tasks stipulated by the law.

National statistical offices across the world already possess an abundance of structured data, both in their databases and files in various formats, but lack the means for exposing, sharing, and interlinking this data on the Semantic Web. Statistical data underpins many of the mash-ups and visualisations we see on the Web, while also being the foundations for policy prediction, planning and adjustments. Making this data available as Linked Open Data would allow for easy enrichment, advanced data manipulation and mashups, as well as effortless publishing and discovery. More specifically, publishing statistical data using RDF comes with the following benefits:

- The individual observations, and groups of observations, become (web) addressable, allowing for third party annotations and linking (e.g. a report can reference the specific figures it is based on, allowing for fine grained provenance trace-back).

[11] http://ec.europa.eu/eurostat

- In RDF, the fact that the data is decoupled from the layout means the layout has no effect on the interpretation of the data, unlike the table approach, where the layout has significant influence on the way the information can be read and interpreted.
- As RDF does not rely on the data being properly laid out, this separation makes it possible to re-contextualize the dataset by embedding or integrating it with another dataset. This further extends to combination between statistical and non-statistical sets within the Linked Data web.
- Datasets can be manipulated in new ways, due to the fine grained representation enabled by the Linked Data approach.
- For publishers who currently only administer static files, Linked Data offers a flexible, non-proprietary, machine readable means of publication that supports an out-of-the-box web API for programmatic access.

4.1 Relevant Standards

We can think of the statistical dataset as a multi-dimensional space, or hypercube, indexed by those dimensions. This space is commonly referred to as a *cube* for short; though the name should not be taken literally, it is not meant to imply that there are exactly three dimensions (there can be more or fewer), nor that all the dimensions are somehow similar in size.

The Statistical Data and Metadata eXchange (SDMX)[12] is an ISO standard used by the U.S. Federal Reserve Board, the European Central Bank, Eurostat, the WHO, the IMF, and the World Bank. The United Nations and the Organization for Economic Cooperation and Development expect the national statistics offices across the world to use SDMX to allow aggregation across national boundaries. However, the fact that the concepts, code lists, datasets, and observations are not named with URIs or routinely exposed to browsers and other web-crawlers, makes SDMX not web-friendly, which, in turn, makes it more difficult for third parties to annotate, reference, and discover that statistical data.

The Data Cube RDF vocabulary[13] is a core foundation, focused purely on the publication of multi-dimensional data on the Web. It supports extension vocabularies to enable publication of other aspects of statistical data flows. As this cube model is very general, Data Cube can also be used for other datasets (e.g. survey data, spreadsheets and OLAP data cubes).

SDMX-RDF is an RDF 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. More specifically, it defines classes and predicates to represent statistical data within RDF, compatible with the SDMX information model.

[12] http://sdmx.org
[13] http://www.w3.org/TR/vocab-data-cube/

4.2 Working with Statistical Linked Data

Within LOD2 a specialized version of the Stack – the *Statistical Workbench* – has been developed to support the need of experts that work with statistical data. Figure 2 shows the main menu of the software environment where the operations are organized into five subgroups called Manage Graph, Find more Data Online, Edit & Transform, Enrich Datacube, and Present & Publish respectively. Once the graph is uploaded into the RDF store, it can be further investigated with the Validation Tool, or with the OntoWiki semantic browser. The validation component checks if the supplied graph is valid according to the integrity constraints defined in the RDF Data Cube specification. Each constraint in the document is expressed as narrative prose, and where possible, SPARQL ASK queries are provided. These queries return true if the graph contains one or more Data Cube instances which violate the corresponding constraint. If the graph contains a well-formed RDF Data Cube, it can be visualized with the CubeViz tool.

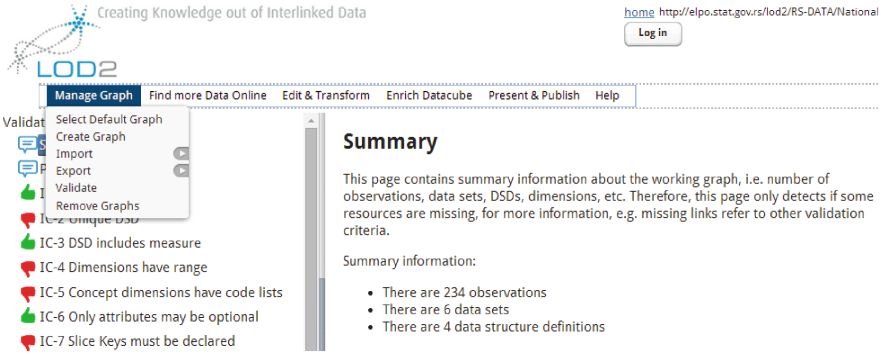

Fig. 2. LOD2 Statistical Workbench

4.3 Serbian Statistical Office Use Case

The information published by the Statistical Office of the Republic of Serbia (SORS)[14] on monthly, quarterly and yearly basis is mostly available as open, downloadable, free of charge documents in PDF format, while raw data with short and long-term indicators is organized in a central statistics publication database. The SORS publication list includes press releases, a monthly statistical bulletin, statistical yearbook, working documents, methodologies and standards, trends, etc. Serbia's national statistical office has shown 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. A number of envisioned main actors,

[14] http://www.stat.gov.rs

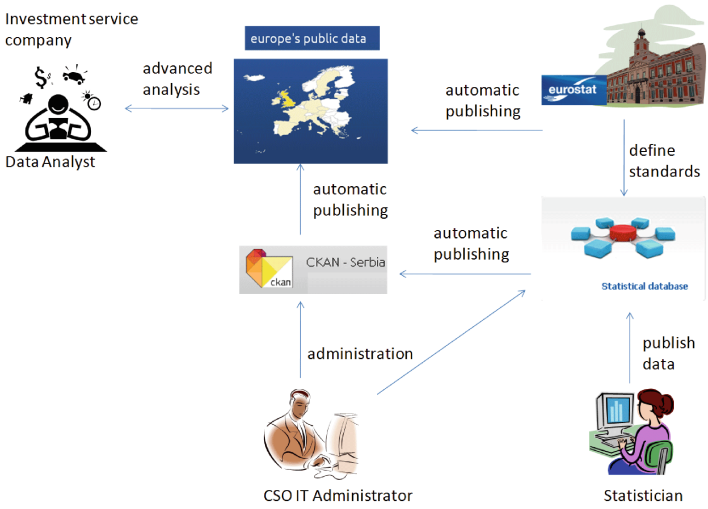

Fig. 3. Simplified workflow for an end-user case

and a sample scenario, were used to elaborate the requirements for the Linked Open Data tools to be included in the Statistical Workbench.

The SORS data publishing process with the Linked Open Data extension is shown in Fig. 3. The data prepared by a Statistician are published through the SORS Dissemination database. Using the LOD2 Statistical Workbench, reports can be transformed into a machine processable format and published to a local governmental portal (e.g. the Serbian CKAN). The IT Administrator maintains the necessary infrastructure for storing and publishing statistical data in different formats (Excel, XML, RDF). Public data are retrieved by a Data analyst that wants to use the data in his research.

An in-depth analysis of the SORS dissemination database has shown that there are a number of standard dimensions that are used to select and retrieve information. The Linked Data principles suggest modeling these dimensions as code lists in accordance with the recommendation for publishing RDF data using the RDF Data Cube vocabulary. In order to formalize the conceptualisation of each of the domains in question, the Simple Knowledge Organisation System (SKOS) was used. The concepts are represented as `skos:Concept` and grouped in concept schemes that serve as code lists (`skos:ConceptScheme`) the dataset dimensions draw on to describe the data (Fig. 4).

As the direct central database access is restricted, all input data is provided as XML files. The SORS statistical data in XML form is passed as input to the Statistical Workbench's built-in XSLT (Extensible Stylesheet Language Transformations) processor and transformed into RDF using the aforementioned vocabularies and concept schemes. Listing 1 shows an example RDF/XML code snippet from a transformed dataset.

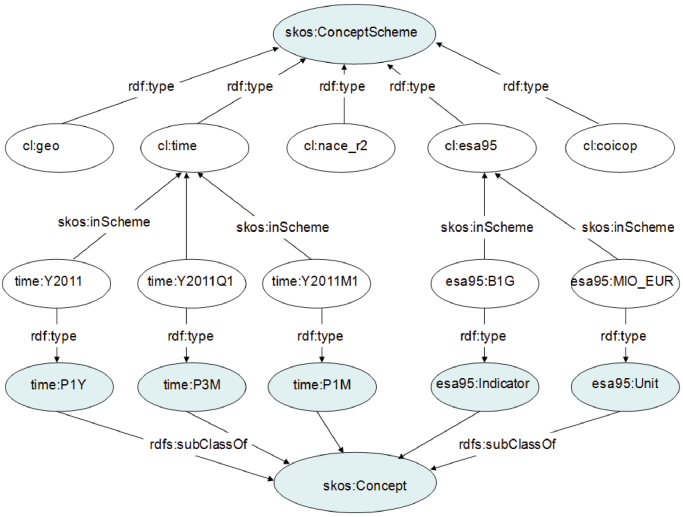

Fig. 4. Representing SORS code lists

Listing 1. Sample observation from a SORS dataset in RDF/XML syntax

```
<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>
```

The above example describes the setup, the overall process and the outcome of the use case. It shows how (local) raw statistical data can be moved to (globally visible) rich collections of interrelated statistical datasets. A future-proof novel method is used for data representation, compatible with international statistical standards. The resulting data relies on both international and domestic code lists, allowing for easy comparison, interlinking, discovery and merging across different datasets. Finally, the results are cataloged in a local metadata repository, and periodical harvesting at an international level is scheduled, thereby increasing transparency and improving public service delivery, while enriching the Linked Data Cloud.

5 Multidimensional Economy Data in Poland

The Polish Ministry of Economy (MoE) publishes various types of yearly, quarterly, monthly and daily economical analyses, based on the data coming from the Ministry of Finance, Central Statistical Office, Warsaw Stock Exchange,

Ministry's own data and other governmental agencies resources. There are also other larger data sets made available either on-line or for download which could be made more easily processable without human intervention. The data as of today is being published in the formats and presentation forms intended for direct consumption by humans, but with very limited possibilities for automated processing or analysis conducted by custom software. Within the LOD2 project the data was lifted to the semantic level.

5.1 Polish Open Economy Data

The primary source of data, subject to publication using the LOD2 tools, was the warehouse for macroeconomic information, namely the INSIGOS database. The data stored by INSIGOS, concerns the Polish economy as well as the international trade of Poland (import and export). The INSIGOS database was built using the internal database of the MoE, namely the Business Information Knowledge Base, maintained since 2001.

The INSIGOS contains the following datasets:

- HZ/GEO – foreign trade broken down by countries
- HZ/CN – foreign trade broken down by type of goods traded, as reported using the Common Nomenclature codes (CN)
- POLGOS – several economic indicators broken down by type of activity, as reported by companies in Poland using F-01 statistical form, using PKD – Polish Classification of Activities. POLGOS was later divided into POLGOS/PKD2004 and POLGOS/PKD2007, because there was no simple way to convert the data to one of the classifications.

5.2 Modelling Multidimensional Data with Data Cube Vocabulary

The most important artefacts that we used from RDF Data Cube Vocabulary include: `qb:DimensionProperty`, `qb:ComponentSpecification`, `qb:DataSet` and `qb:DataStructureDefinition`. The model prepared for INSIGOS HZ module is presented in Fig. 5.

In INSIGOS HZ module there are only two indicators: *import* and *export*. SKOS concept scheme was created to facilitate the use of measures in dimensions and then in observations. The name of the `skos:ConceptScheme` is `HZ` and it was assigned by `qb:CodedProperty`. Both measure properties were defined according to pattern presented in Listing 2.

Listing 2. Sample measure property from INSIGOS HZ/GEO dataset in turtle

```
<http://data.i2g.pl/insigos/properties/import>
  a qb:MeasureProperty, skos:Concept ;
  skos:prefLabel "Import"@pl ; skos:prefLabel "Import"@en ;
  rdfs:range xsd:decimal ;
  dcterms:source <https://insigos.mg.gov.pl/DaneHZFiltr.aspx> ;
  dcterms:publisher "Polish␣Ministry␣of␣Economy"@en ;
  dcterms:publisher "Ministerstwo␣Gospodarki"@pl .
```

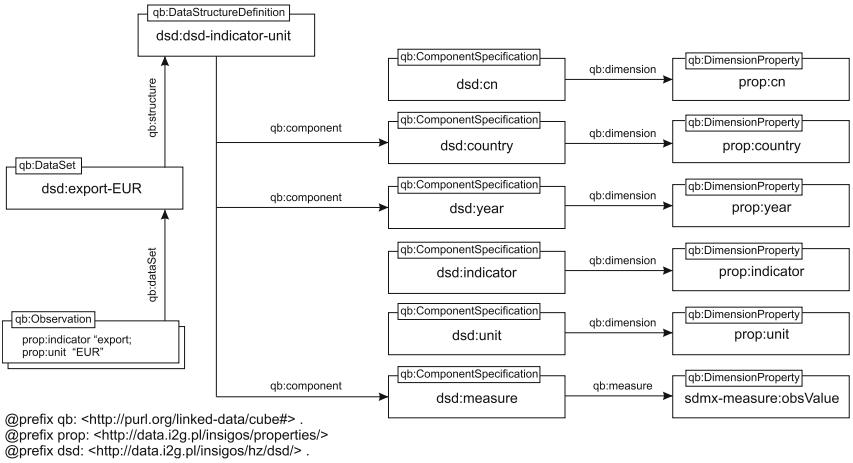

Fig. 5. INSIGOS HZ modelled in Data Cube vocabulary

5.3 Slices Generation

After conversion of source data to data cube we still have just detailed data. In order to show some useful statistics, and in particular to visualise in CubeViz, we need to provide slices and aggregations. Such artefacts need to be modelled and included along with RDF data.

In order to generate slices we provided a Python script for materialisation of datasets. It first queries for a list of dimensions enumerated in data structure definition (DSD). Then, all elements (members) of dimensions are retrieved. We assume that slice contains two dimensions (for use in two-dimensional charts), therefore pairwise combination of all dimensions is generated. Finally, respective links between observations, datasets and slices are generated.[15]

Listing 3 presents relevant parts of our generated data. In this example, a *slicekey* with fixed dimensions for indicator and unit is defined – `slicekey-indicator-unit`. Using this structure, several slices are generated, one for each combination of free variables. One of them is the slice with values fixed on 'export' (*indicator*) and 'EUR' (*unit*) – `export-EUR`. The last line by `qb:observation` contains ellipsis because there are in fact 1330 observations attached.

[15] A mechanism should be introduced to allow querying for observations instead of explicit assignments. Currently, such assignments require materialisation of a big number of additional triples, which makes solution questionable in enterprise data warehouse settings when considering the volume of data.

Listing 3. Data cube slices in INSIGOS data

```
@prefix dsd: <http://data.i2g.pl/insigos/hz/dsd/> .
dsd:dsd a qb:DataStructureDefinition ;
  rdfs:label "Complete␣DSD␣for␣indicator-unit"@en ;
  qb:component dsd:year, dsd:unit, dsd:indicator, dsd:country,
    dsd:measure ;
  qb:sliceKey dsd:slicekey-indicator-unit .
dsd:slicekey-indicator-unit a qb:SliceKey ;
  rdfs:label "slicekey␣by␣indicator-unit"@en ;
  qb:componentProperty
    <http://data.i2g.pl/insigos/properties/year>,
    <http://data.i2g.pl/insigos/properties/country> .
dsd:dataset a qb:DataSet ;
  rdfs:label "Dataset␣with␣all␣dimensions"@en ;
  qb:structure dsd:dsd;
  qb:slice dsd:slice-export-EUR .
dsd:slice-export-EUR a qb:Slice ;
  rdfs:label "export␣-␣EUR" ;
  qb:sliceStructure dsd:slicekey-indicator-unit ;
  qb:observation
    <http://data.i2g.pl/insigos/hz/geo/export/AD/2006/EUR>, ...
```

5.4 Aggregations

Aggregations are commonly used in reporting using multidimensional data. The most prominent example is a data warehouse with OLAP cubes. Aggregations are selected and calculated in such a way that speeded up reporting is possible. By analogy, aggregations may be deemed also useful for cubes defined as linked data.

In our case aggregation is necessary for drill-down operations. For example, daily data can be aggregated on a monthly basis to better observe a phenomenon. Also, we can display data in yearly sums, and allow drill-down to be even more precise. SPARQL is capable of calculating sums on-the-fly, but it takes time and sometimes time-out is reached.[16] Materialisation is then necessary for quicker operations.

Our first idea was to prepare aggregations using Python script, similar to slicing. That would require too much querying and would be inefficient. In the end, we found a way to implement the method for aggregation as a set of SPARQL queries.

One of the issues was generation of URIs for new observations as aggregation is in fact a new observation – the same dimensions but values are on higher level, e.g. month → year. For INSIGOS/POLGOS observations we have defined a pattern for identifiers. We used the capabilities of Virtuoso to generate identifiers directly in SPARQL.

[16] For example, a query calculating the value of public procurement contracts by voivodeships takes 100 seconds, which is outside of acceptable response times.

Before aggregation is done, a correct hierarchy should be prepared. The prerequisite for the script is that dimensions are represented as SKOS concept scheme, and elements of dimension are organised in hierarchy with *skos:narrower* property.

6 Exploration and Visualisation of Converted Data

In the following we describe two scenarios to showcase benefits of the presented framework. The first one is about statistical data discovery using CubeViz. The second scenario is about discovering geospatial information by the use of Facete.

6.1 Statistical Data Exploration

CubeViz, the RDF Data Cube browser, depicted in Fig. 6 allows to explore data described by RDF Data Cube vocabulary [2]. CubeViz generates facets according to the RDF Data Cube vocabulary artefacts such as Data Cube DataSet, Data Cube Slice, a specific measure and attribute (unit) property and a set of dimension elements that are part of the dimensions.

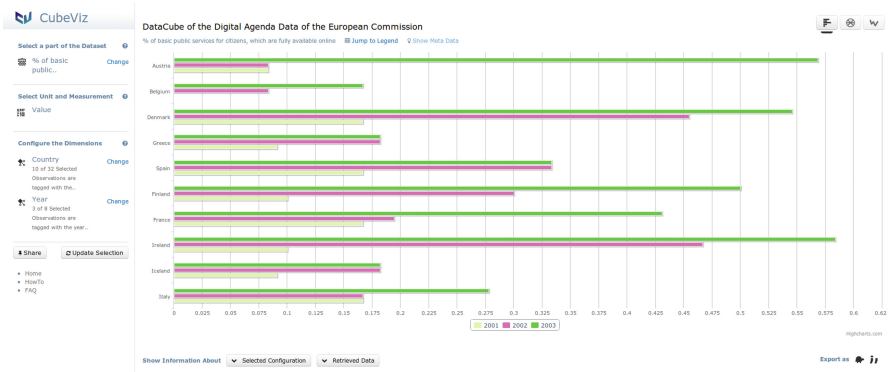

Fig. 6. Screenshot of CubeViz with faceted data selection and chart visualization component.

Based on the selected facets, CubeViz retrieves data from a triplestore and suggests possible visualizations to the user. Users or domain experts are able to select different types of charts such as a bar chart, pie chart, line chart and polar chart that are offered depending on the selected amount of dimensions and respective elements.

6.2 Geospatial Data Discovery

Facete, depicted in Fig. 7, is a novel web application for generic faceted browsing of data that is accessible via SPARQL endpoints. Users are empowered to create

custom data tables from a set of resources by linking their (possibly nested) properties to table columns. A faceted filtering component allows one to restrict the resources to only those that match the desired constraints, effectively filtering the rows of the corresponding data table. Facete is capable of detecting sequences of properties connecting the customized set of resources with those that are suitable for map display, and will automatically show markers for the shortest connection it found on the map, while offering all further connections in a drop down list. Facete demonstrates, that meaningful exploration of a spatial dataset can be achieved by merely passing the URL of a SPARQL service to a suitable web application, thus clearly highlighting the benefit of the RDF transformation.

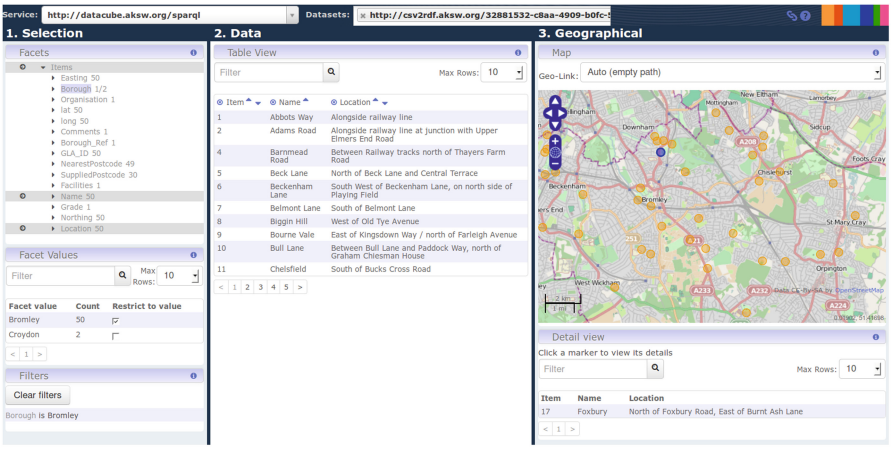

Fig. 7. Screenshot of Facete showing data about allotments in South East London.

6.3 Drill-Down Choropleth Maps

Import and export statistics of Poland collected in INSIGOS HZ/GEO dataset are best visualised on the globe. The globe itself is part of D3 library[17]. Some work was, however, necessary in order to allow display of data from triple store. Several parameters are defined in the graphical interface, and based on it SPARQL queries are prepared. Then, the legend is defined in such a way that colours are more or less equally distributed. Normally the numbers for import and export are subject to power law, therefore the legend scale cannot be linear. The map is coloured according to values assigned to selected countries. A sample map is presented in Fig. 8 shows 20 countries with the greatest value of export in 2012 expressed in PLN (Polish currency), with the unit being millions.

Not only technical communication with Virtuoso had to be solved. We first needed to integrate data on semantic level, i.e. map of the world in D3 had

[17] Data-Driven Documents, D3.js, http://d3js.org/.

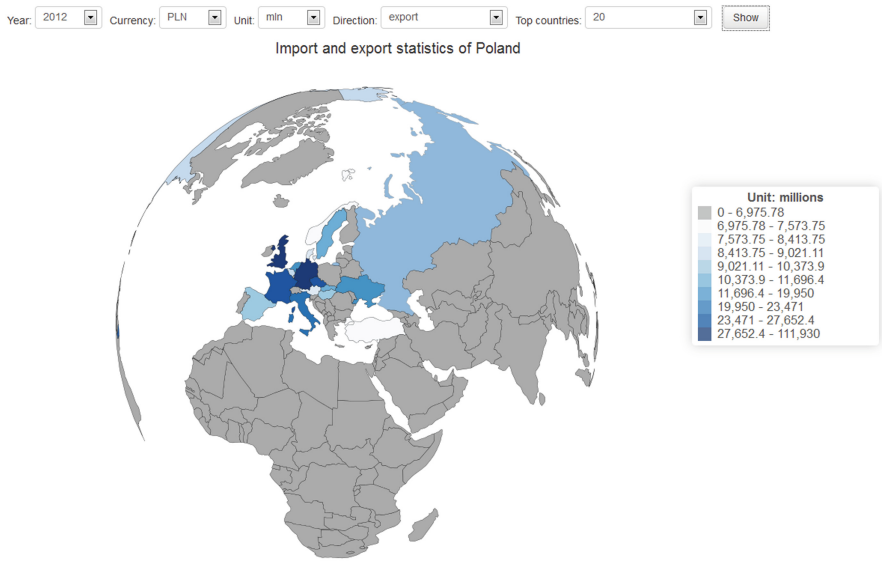

Fig. 8. Export statistics of Poland in 2012 presented on a globe

country codes consisting of three letters. Countries in INSIGOS dataset had just names, and therefore additional mapping was necessary. It should be noted that list of countries changed for the period analysed but the map has not been updated.

More popular is visualisation of data on the country level. For this purpose we need a map of country available in SVG and compatible with D3 library. It can also be derived from open data. For example, a map of Poland has been prepared based on OpenStreetMap and administrative division included there.

On the map of Poland we visualise data concerning public contracts. Several measures can be visualised like: number of contractors, number of contracts or value of contracts. All measures are sliced by geographical dimensions reflecting administrative division of Poland. There are 16 voivodeships (*województwo*) and 380 districts (*powiat*) in Poland. Showing 380 entities on the map is not very useful for interpretation. Therefore we have applied the drill-down approach. First, a user is presented with a map of the whole Poland with 16 voivodeships. Then, after clicking on selected region, the user goes to detailed map of districts in a given region. There is also another administrative level – county (*gmina*) – which can be included when needed. Analogous maps can be prepared for other countries as well.

6.4 Drill-Down Tables

In terms of the drill-down functionality, we need to remember that datasets can be aggregated on various level of detail and very often they are offered as the

same package. Geography is not the only dimension. There are several others that cannot be visualised within the map, hence the need to develop a drill-down table. Some examples include (in the case of Polish data): time dimension; Polish classification of activities (PKD, NACE): sections and chapters; common nomenclature (CN): several levels; various economic indicators in energy-related data (e.g. production total → from renewable sources → from solar plants).

Due to required integration with triple store we prepared our own drill-down table from scratch. The prerequisite is that the dimension to be used for drill-down is described as SKOS concept scheme. It is an industry standard and allows to represent hierarchies conveniently. It has also mechanism for labels in various languages. Alternative labels make mapping to this headers more flexible when heterogeneous sources are considered. All vocabularies, including time dimension, were prepared with this prerequisite in mind.

There are in fact three queries necessary to prepare a drill-down table. The approach is thus similar to multidimensional queries against OLAP cubes in MDX[18]. First, we need to get headers of rows and columns, then data itself. Not only labels for headers are necessary but also interdependencies between headers and their level. When a drill-down table is first loaded, rows and columns are collapsed, so that only most aggregated data is shown. It is then possible to click on the respective row or column marked with 'plus' sign to expand one level. Figure 9 presents expanded columns.

Year	⊟ Income	⊟ Production altogether	⊞ Proffesional power plants	Industrial thermal-electric power stations	⊞ Independent power plants	Import (consumption)	⊞ Expenditure
⊞ 2007	167109	159348	148977	7666	2705	7761	167109
⊞ 2008	163975	155494	144774	7937	2783	8481	163975
⊞ 2009	158612,778204	151210,163204	140312,696	6560,414	4337,053204	7402,615	158612,778204
⊞ 2010	161332,618	155038,723	143294,293	7267,168	4477,262	6293,895	161332,618

Fig. 9. Drill-down table with expanded columns

7 Conclusions and Future Work

Opening up governmental data requires two elements: discoverability – where more data portals are available, harvesting is used to gather data; and quality machine readable data – where data can be trusted. LOD2 tools support development of these functionalities.

We have addressed the general challenge where data currently being published is in formats and presentation forms intended for direct consumption by humans, but with limited possibilities for automated processing or analysis

[18] MultiDimensional eXpressions, a query and manipulation language for OLAP databases.

conducted by a custom software. As one of the important issues for data discoverability we have identified the need for providing data catalogues. CKAN is a reference solution in this area but it requires further extensions. Working implementations include: http://publicdata.eu and http://open-data.europa.eu. The quality of data has been demonstrated by providing Polish and Serbian data first automatically converted and then carefully improved.

One of the missing features is cross-language searching. Although the user interface can be used in multiple languages, metadata for datasets can of course be read only in the language in which it was input. A limitation of search on PublicData.eu is that as the source catalogues are naturally in different languages, a single search term in any given language will usually not find all the relevant datasets.

For wider adoption of CKAN we also need better metadata management. The current harvesting arrangement does not preserve all the original metadata. In particular, even where the harvest source is CKAN-based and datasets have identifiable publishing departments, this information is not preserved in the record on PublicData.eu. Adding this information to the harvesting system would enable users on PublicData.eu to 'follow' individual departments and see dashboard notifications when departments published new or updated data. An example of an alternative harvesting process (built on top of LOD2 stack tools) that preserves the original metadata is available at http://data.opendatasupport.eu.

Several other tools, not mentioned in the chapter, have been particularly useful for making data accessible for machines: Virtuoso, Ontowiki (with CubeViz plug-in), SILK, Open Refine and PoolParty. Various datasets have been elaborated in detail manually, particularly those using the RDF Data Cube vocabulary. Some examples include: national accounts, foreign trade, energy-related, and public procurement data. We have increased the openness of the data by preparing respective vocabularies and providing linking to other data sources available on the Web.

A significant amount of time was absorbed by data quality issues. Even though data was available in 'machine processable' XML, it were users who entered incorrect data. These are, however, typical problems of integration projects and should not under any circumstances be considered to be related to the linked data paradigm. On the contrary, applying tools that we had at our disposal allowed to spot quality problems even faster than we would have been able to otherwise.

References

1. Berners-Lee, T.: Relational databases on the semantic web, 09 1998. Design Issues. http://www.w3.org/DesignIssues/RDB-RDF.html
2. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube vocabulary. Technical report, W3C (2013)
3. Ermilov, I., Auer, S., Stadler, C.: Csv2rdf: user-driven csv to rdf mass conversion framework. In: ISEM '13, 04–06 September 2013, Graz, Austria (2013)
4. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. J. Web Semant. **5**, 251–261 (2007)