CSV2RDF: User-Driven CSV to RDF Mass Conversion Framework

Ivan Ermilov[1], Claus Stadler[1], Michael Martin[1], and Sören Auer[2]

[1] AKSW/BIS, Universität Leipzig, PO BOX 100920, 04009 Leipzig, Germany
firstname.lastname@informatik.uni-leipzig.de
[2] CS/EIS, Universität Bonn, Römerstraße 164, 53117 Bonn, Germany
auer@cs.uni-bonn.de

Abstract. Governments and public administrations started recently to publish large amounts of structured data on the Web, mostly in the form of tabular data such as CSV files or Excel sheets. We showcase an application for user-driven transformation and visualization of tabular data from data portals such as PublicData.eu to RDF. It 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 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 mining tasks on the transformed tabular data. The application of our approach to the PublicData.eu portal results in 10.000 transformed datasets amounting 7.3 Billion triples, thus adding a sizable part to the Web of Data.

1 Introduction

Integrating and analyzing 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. Despite recent advances in structured data publishing on the Web (such as RDFa and the schema.org initiative) the question arises how larger datasets can be published, described in order to make them easily discoverable and facilitate the integration as well as analysis.

One approach for addressing this problem are data portals, which enable organizations to upload and describe datasets using comprehensive metadata schemes. Similar to digital libraries, networks of such data catalogs can support the description, archiving and discovery of datasets on the Web. Recently, we have seen a rapid growth of data catalogs being made available on the Web. The data catalog registry *datacatalogs.org*, for example, lists already 362 data catalogs worldwide. Examples for the increasing popularity of data catalogs are Open Government Data portals, data portals of international organizations and NGOs as well as scientific data portals.

Governments and public administrations started to publish large amounts of structured data on the Web, mostly in the form of tabular data such as CSV files or Excel sheets. Examples are the data portals of the US, the UK or the European Commission as well as numerous other local, regional and national data portal initiatives.

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 unquestioned 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, for example, only 459 out of more than 17.000 datasets (i.e. just 3%) 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.

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.

2 PublicData.eu Data Overview

PublicData.eu is a data catalog aiming to become a one stop shop for open-data in Europe. The rationale is to increase public access to high-value, machine-readable datasets generated by the European, national, regional as well as local governments and public administrations. This is achieved by harvesting and exposing datasets from various European data catalogs (currently 17 catalogs are harvested[3]).

At the time of writing PublicData.eu comprises 20,396 datasets. Each dataset can comprise several data resources and there are overall 55,000+ 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 [1].

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

[3] `http://www.datacatalogs.org/dataset?groups=publicdata-eu`

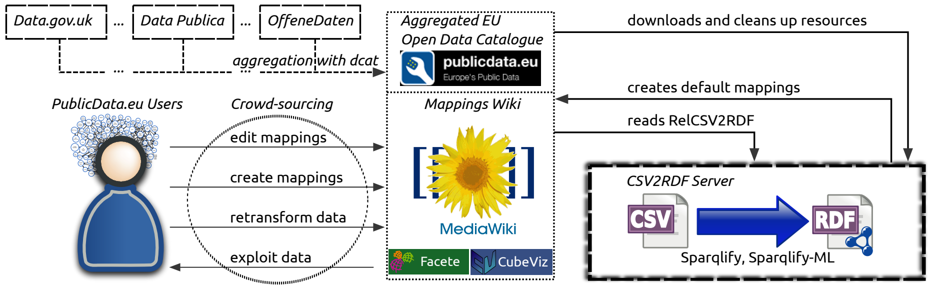


Fig. 1: Architecture of our CSV2RDF extension for PublicData.eu.

preserve much of the domain semantics and structure. Also, tabular data represented in the above mentioned formats can be syntactically quite heterogeneous[4] 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.

3 User-driven Conversion Framework

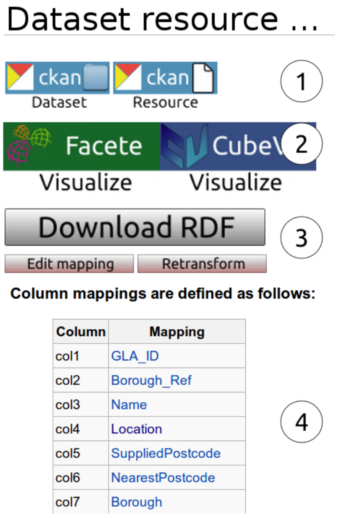
The completely automatic RDF transformation as well as the detection and correction of tabular data problems is not feasible. Therefore, we devise an approach where the effort is shared between machines and human users. Our mapping authoring environment is based on the popular *MediaWiki*[5] system. 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. In order to realize the automatic conversion, our implementation downloads and cleans resources available on PublicData.eu. In a next step it extracts the header of the tabular data file, creates a default mapping automatically and converts the data based on this mapping to RDF using Sparqlify-CSV as described in the previous section. Finally, a page on wiki.publicdata.eu is created for each resource containing the mappings, links to rerun the transformation routine, download links for the resulting RDF files as well as links to CubeViz and Facete visualizations. An overview of the entire application is depicted in Figure 1.

Our application continuously crawls CSV resources from PublicData.eu and validates them. Around 20% of CSV resources are filtered out, mostly because of

[4] Informational RFC for CSV: `http://www.ietf.org/rfc/rfc4180.txt`

[5] `http://www.mediawiki.org/`



Column mappings are defined as follows:

Column	Mapping
col1	GLA_ID
col2	Borough_Ref
col3	Name
col4	Location
col5	SuppliedPostcode
col6	NearestPostcode
col7	Borough

```
{{CSV2RDFHeader}}

...

{{RelCSV2RDF
 | name     = default-mapping
 | header   =  1
 | omitRows = -1
 | omitCols = -1
 | delimiter =
 | col1 = Department Family
 | col2 = Entity
 | col3 = Payment Date
 | col4 = Expense Type
 | col5 = Cost Centre Name
 | col6 = Supplier
 | col7 = Transaction No.
 | col8 = Line Amount
 | col9 = Invoice Total
}}
```

Fig. 2: Dataset resource page on wiki.publicdata.eu with mapping definition (left): 1 - links to the CKAN package and resource descriptions; 2 - links to visualizations by CubeViz and Facete; 3 - download, edit and retransform buttons; 4 - transformation mapping. The wiki text mark up for the mapping (right).

response timeouts, server errors or missing files. The second step after validation is the automatic creation of the default mapping and the conversion to RDF. In order to obtain an RDF graph from a table T we essentially use the *table as class* approach [2], 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 name space. 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 yet too complicated for novice users and therefore less suitable for being crowd-sourced. To even 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* (e.g. Figure 2). The complete description for the template is available on the mapping wiki[6].

At the end of the transformation a page is created for each resource on the mappings wiki at wiki.publicdata.eu (e.g. Figure 2). 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

[6] http://wiki.publicdata.eu/wiki/Mapping_syntax

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

Table 1: Transformation results summary.

RelCSV2RDF template into a human-readable description of the parameters including links for transformation rerun and RDF download.

The mapping wiki uses the *Semantic MediaWiki* [3] (SMW) extension, which enables semantic annotations and embedding of search queries over these annotation 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, that is dataset resources are connected through the properties used in their mappings.

4 Results

We downloaded and cleaned 9,370 CSV files, that consume in total 33 GB of disk space. The distribution of the file sizes shows, that the vast majority (i.e. 85%) of the published datasets are less than 100 kB in the size. 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.

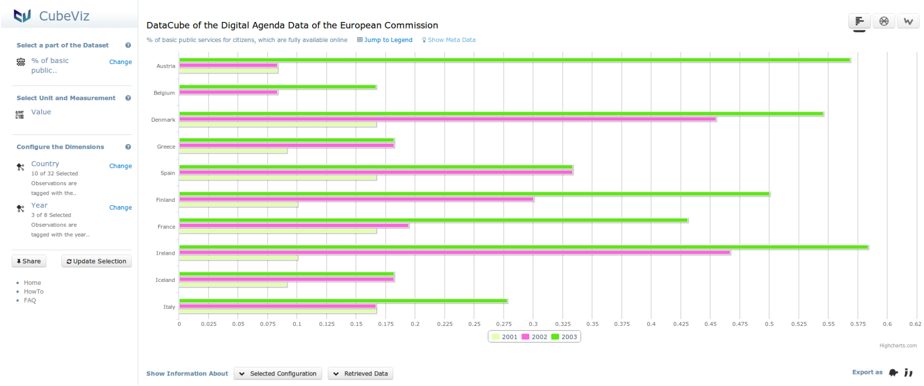
During the automatic conversion our framework created 9,370 wiki pages on the mappings wiki. The `has_property` property is used 80,676 times and maps to 13,490 distinct properties. The 3 most used properties are: Entity (occurs in 3,593 resources), Supplier (3,505) and Amount (3,151). The full list of most used properties is located on the mapping wiki[7].

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 50MB 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 50MB, the transformation is currently queued and processed in batch mode.

5 Discovery 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

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



CubeViz

Select a part of the Dataset

% of basic public.. Change

Select Unit and Measurement

Value

Configure the Dimensions

Country Change
10 of 32 Selected
Observations are tagged with the..

Year Change
3 of 8 Selected
Observations are tagged with the year..

DataCube of the Digital Agenda Data of the European Commission

% of basic public services for citizens, which are fully available online

Fig. 3: Screenshot of CubeViz with facetted data selection and chart visualization component.

described in subsection 5.1. The second scenario is about discovering geospatial information by the use of Facete presented in section subsection 5.2.

5.1 Statistical Data Discovery

Once statistical data is represented in a tabular form was converted to the RDF DataCube vocabulary [4] using the CSV2RDF converter[8] (usage presented in screencast[9]), the user is able to discover the data by using CubeViz, the RDF DataCube browser[10]. CubeViz generates facets as illustrated in Figure 3 according to the RDF DataCube vocabulary such as follows:

1. Selection of a DataCube DataSet,
2. Selection of a DataCube Slice,
3. Selection of a specific measure and attribute (unit) property and
4. Selection of a set of dimension elements that are part of the dimensions.

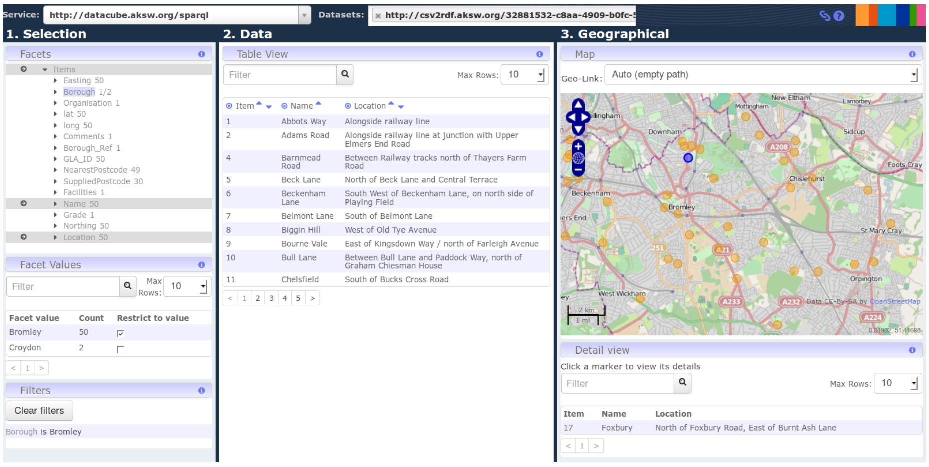
After finalizing the selection using those facets, a SPARQL query will be generated in order to retrieve all matching observations. Afterwards, the result set is analyzed to detect the amount of dimensions containing multiple elements and to select the charts that can be used to visualize the selected observation. As an outcome of the analysis, the first entry from the chart list will be selected and the conditioned result set is assigned to it. Further configurations adjustable in CubeViz act on the visualization level. 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 its respective elements.

After rendering a chart, CubeViz offers chart-specific options, that can be used to adjust the output according to the users needs. For instance, in order

[8] https://github.com/AKSW/csvimport.ontowiki

[9] http://www.youtube.com/watch?v=Ib8b8YWU2i8

[10] http://aksw.org/Projects/CubeViz



Service: http://datacube.aksw.org/sparql

Datasets: http://csv2rdf.aksw.org/32881532-c8aa-4909-b0fc-5

1. Selection

Facets

- Items
  - Easting 50
  - Borough 1/2
  - Organisation 1
  - lat 50
  - long 50
  - Comments 1
  - Borough_Ref 1
  - GLA_ID 50
  - NearestPostcode 49
  - SuppliedPostcode 30
  - Facilities 1
  - Name 50
  - Grade 1
  - Northing 50
  - Location 50

Facet Values

Max Rows: 10

Facet value	Count	Restrict to value
Bromley	50	☑
Croydon	2	□

Filters

Borough is Bromley

2. Data

Table View

Max Rows: 10

Item	Name	Location
1	Abbots Way	Alongside railway line
2	Adams Road	Alongside railway line at junction with Upper Elmers End Road
4	Barnmead Road	Between Railway tracks north of Thayers Farm Road
5	Beck Lane	North of Beck Lane and Central Terrace
6	Beckenham Lane	South West of Beckenham Lane, on north side of Playing Field
7	Belmont Lane	South of Belmont Lane
8	Biggin Hill	West of Old Tye Avenue
9	Bourne Vale	East of Kingsdown Way / north of Farleigh Avenue
10	Bull Lane	Between Bull Lane and Paddock Way, north of Graham Chiesman House
11	Chelsfield	South of Bucks Cross Road

3. Geographical

Map

Geo-Link: Auto (empty path)

Detail view

Click a marker to view its details

Max Rows: 10

Item	Name	Location
17	Foxbury	North of Foxbury Road, East of Burnt Ash Lane

Fig. 4: Screenshot of Facete showing data about allotments in South East London.

to display widespread measurement values a logarithmic scale can be selected for improved visualization experience. Further integrated adjustment options are the chart subtype (offering combinations, e.g. polar/column chart) and the switch/inversion of the axis and dimensions. After configuring the chart, it is possible to share it within a community using the permanent link or exports in CSV or RDF-Turtle notation.

5.2 Geospatial Data Discovery

Facete, depicted in Figure 4, is a novel web application for generic faceted browsing of data that is accessible via SPARQL endpoints.[11] 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.

References

1. Ermilov, I., Auer, S., Stadler, C.: Csv2rdf: User-driven csv to rdf mass conversion framework. In: ISEM '13, September 04 - 06 2013, Graz, Austria. (2013)

[11] https://github.com/GeoKnow/Facete, Screencast: http://www.youtube.com/watch?v=VzEvFJs89Wc

2. Berners-Lee, T.: Relational databases on the semantic web. (09 1998) Design Issues, http://www.w3.org/DesignIssues/RDB-RDF.html.
3. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic wikipedia. Journal of Web Semantics **5** (September 2007) 251–261
4. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube vocabulary. Technical report, W3C (2013) http://www.w3.org/TR/vocab-data-cube/.

Appendix: Addressing Big Data Track Requirements

5.3 Minimal Requirements

1. *Data Volume.* The CSV2RDF applications makes use of a large quantity of partially very large datasets amounting overall 7.3B triples.
2. *Data Variety.* The data includes almost 10,000 resources from hundreds of data publishers in a large variety of different CSV structures from across the European Union. For mapping and enriching this data we implemented a collaboration environment employing a semantic wiki, where the effort is shared between humans and machines.
3. *Data Velocity.* The datasets processed by CSV2RDF and obtained from PublicData.eu are consistently updated and crawled.

5.4 Additional Desirable Features

- *The application should do more than simply store/retrieve large numbers of triples.* In addition to storing and retrieving large numbers of triples, CSV2RDF allows map tabular data to RDF, crowd-source the mapping proces and visualize the results. It also provides tools to perform data mining on the extracted statistical and geospatial data. A search and navigation through the metadata of the converted datasets is also available.
- *The application should be scalable in terms of the amount of data used.* CSV2RDF has been proven to scale in terms of the number of datasets to be processed, individual dataset size and with regard to the number of mapping users.
- *The application should be scalable in terms of distributed components working together.* Through PublicData.eu CSV2RDF obtains dataset metadata from more than 30 different data catalogs. The datasets itself are downloaded from hundreds of individual data publishers sites.
- *The application should either function in real-time or have a real-time realization.* CSV2RDF continuously crawls and transforms datasets registered at one of the data catalogs aggregated via PublicData.eu. The mapping transformation can be executed in real-time for all datasets smaller than 100MB (i.e. 99% of the datasets). The exploration and visualization interfaces again can be configured and used in real-time.