## Deliberable 9.1.3: Final Release of the Publicdata.eu Website and Tools

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date of Deliverable</td>
<td>M48, 31/08/14</td>
</tr>
<tr>
<td>Actual Submission Date</td>
<td>M48, 31/08/14</td>
</tr>
<tr>
<td>Work Package</td>
<td>WP 9, Use Case 3: LOD2 for Citizens: Publicdata.eu</td>
</tr>
<tr>
<td>Task</td>
<td>T 9.1</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Approval Status</td>
<td>Approved</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of Pages</td>
<td>23</td>
</tr>
<tr>
<td>Filename</td>
<td>lod2-d9-1-3-publicdata-eu.pdf</td>
</tr>
</tbody>
</table>

**Abstract:** This document provides a summary of the work and results on the Publicdata.eu website and related work.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.
### History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason</th>
<th>Revised by</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>21/08/2014</td>
<td>First draft</td>
<td>Sander van der Waal (OKFN)</td>
</tr>
<tr>
<td>0.2</td>
<td>26/08/2014</td>
<td>First full version submitted for review</td>
<td>Sander van der Waal (OKFN)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ivan Ermilov (ULEI)</td>
</tr>
<tr>
<td>1.0</td>
<td>01/09/2014</td>
<td>Updated to reflect reviewer’s comments</td>
<td>Sander van der Waal (OKFN)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ivan Ermilov (ULEI)</td>
</tr>
</tbody>
</table>

### Author List

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Name</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>OKFN</td>
<td>Sander van der Waal</td>
<td><a href="mailto:sander.vanderwaal@okfn.org">sander.vanderwaal@okfn.org</a></td>
</tr>
<tr>
<td>ULEI</td>
<td>Ivan Ermilov</td>
<td><a href="mailto:iermilov@informatik.uni-leipzig.de">iermilov@informatik.uni-leipzig.de</a></td>
</tr>
</tbody>
</table>
# Table of Contents

1. **INTRODUCTION AND OVERVIEW**.................................................................5  
   1.1 THE NEED FOR AN INTEGRATED PAN-EUROPEAN OPEN DATA PORTAL........5  

2. **DELIVERY TIMELINE**.....................................................................................7  
   2.1 EARLY RELEASES.........................................................................................7  
   2.2 WORK DONE IN SEPT 2013 – AUG 2014 ...................................................7  

3. **TECHNOLOGY**..............................................................................................8  
   3.1 CKAN...........................................................................................................8  
   3.2 ADDITIONAL SOFTWARE AND TOOLS.......................................................8  
      *CKAN DCAT extension*...............................................................................8  
      *Recline Data Explorer (http://reclinejs.com/)*.........................................8  
      *Custom-built datacatalogs.org*....................................................................9  

4. **HARVESTED CATALOGUES**.........................................................................9  
   4.1 DATACATALOGS.ORG..................................................................................9  

5. **SUSTAINABILITY VIA CKAN ASSOCIATION**..............................................10  

6. **KEY STATISTICS ON PUBLICDATA.EU**....................................................11  

7. **DATA LIFTING AND HARMONIZATION**....................................................14  

8. **FUTURE OF PUBLICDATA.EU**......................................................................23
List of Figures

Figure 1 - New front page of datacatalogs.org, as released in August 2014 .......................... 10
Figure 2 - Visits on publicdata.eu since the portal was first launched........................................ 12
Figure 3 - Number of visits per day ......................................................................................... 12
Figure 4 - The CSV2RDF application architecture overview .................................................... 14
Figure 5 - Publicdata.eu portal. Dataset view. Navigation to the Mappings Wiki ........................ 15
Figure 6 - Publicdata.eu portal. Resource view. Navigation to the Mappings Wiki ..................... 15
Figure 7 – Publicdata.eu portal. Data view. Example of object-per-row data .............................. 16
Figure 8 - The Mappings Wiki. Mapping example with wiki markdown view .......................... 17
Figure 9 - The Mappings Wiki. GUI mappings editor ................................................................. 18
Figure 10 - Sgvizler visualization over the integrated RDF data for “2010 September Return” ...... 20
Figure 11 - Screenshot of Facete showing data about allotments in South East London .......... 21
Figure 12 - Publicdata.eu portal. Data view. Example of object-per-cell data ............................ 21
Figure 13 - Screenshot of CubeViz with facetted data selection and chart visualization component. 22
1. Introduction and overview

Publicdata.eu is the prototype of a pan-European data catalogue and federation mechanism, developed by the Open Knowledge Foundation as part of the FP7-funded LOD2 project. Based on the CKAN\(^1\) open-source data portal, the prototype site has been developed as a use case and an early adopter of the LOD2 linked data stack technologies. Changes to the CKAN code base made for publicdata.eu have been synchronised with the main CKAN software development and releases, which also include work invested by the Open Knowledge Foundation outside of LOD2.

This Deliverable summarises development around publicdata.eu between September 2013 and August 2014, which marks the end of the LOD2 project. It includes discussion about publicdata.eu as well as developments around the sustainability of CKAN, the open source software that underpins publicdata.eu and which has been co-developed as part of LOD2. We also discuss the data lifting and harmonization work, which has been led by the University of Leipzig.

1.1 The need for an integrated pan-European open data portal

During the lifespan of LOD2 there has been a big increase of interest in opening up official information for the public to reuse. At a national level, numerous member states have national data catalogues, from the Digitaliser.dk data portal run by the Danish National IT and Telecom Agency\(^2\) to the UK’s data.gov.uk site, launched under the direction of the founder of the World Wide Web, Sir Tim Berners-Lee\(^3\). There are countless city level initiatives across Europe as well -- from Helsinki to Munich, Paris to Zaragoza, with many planning to launch soon.

This deluge of open, freely reusable data creates significant social and economic opportunities for European citizens. Transparency in government finances allows journalists and NGOs to hold governments to account based on actual figures. New digital services enable users to have information they are interested in delivered directly to them via email, web or mobile updates. For example, in the UK, TheyWorkForYou lets users know every time their elected representative speaks or when a topic of interest to them is discussed in the British Parliament\(^4\). Also, large complex datasets can be broken down and presented in more intuitive ways. For example the WhereDoesMyMoneyGo project allows users to see where their taxes go using simple and intuitive data visualisation technologies in order to demystify a complex subject\(^5\).

Efforts are underway to link and combine datasets from a large number of different sources. This newfound data integration will ultimately allow developers to create new digital services capable of dealing with increasingly sophisticated questions and queries\(^6\).

In addition to increasing transparency and improving public service delivery, open data creates opportunities for businesses to build new kinds of commercial services around this new data. This opportunity is possible both because they have access technically to the data and because the data is legally accessible to use and reuse. A recent study estimates the market based on European public sector information could be worth as much as €27 billion\(^7\).

\(^1\) http://ckan.org/
\(^2\) http://bit.ly/dk-cat
\(^3\) http://data.gov.uk/
\(^4\) http://www.theyworkforyou.com/
\(^5\) http://wheredoesmymoneygo.org/
\(^6\) http://linkeddata.org/
\(^7\) Estimate from the 2006 MEPSIR study, commissioned by the European Commission. See: http://bit.ly/mepsir
In order to unlock the potential of digital public sector information, developers and other prospective users must be able to find datasets they are interested in reusing. Publicdata.eu provides a single point of access to open, freely reusable datasets from numerous national, regional and local public bodies throughout Europe. Information about European public datasets is currently scattered across many different data catalogues, portals and websites in many different languages, implemented using many different technologies. The kinds of information stored about public datasets may vary from country to country, and from registry to registry. Publicdata.eu harvests and federates this information to enable users to search, query, process, cache and perform other automated tasks on the data from a single place. This helps to solve the “discoverability problem” of finding interesting data across many different government websites, at many different levels of government, and across the many governments in Europe.
2. Delivery Timeline

Publicdata.eu has been developed in an agile fashion, releasing early and often to the main site to maximise user engagement. This has allowed us to go through several iterations of the site and make numerous improvements whilst maximizing value. Over the last year, focus has moved away from developing new features but instead focused more on the longer term sustainability of the CKAN codebase.

Publicdata.eu is based on CKAN and many of the features that have been developed for Publicdata.eu have been contributed back to CKAN. As such, the development of CKAN has benefited immensely from the support of the LOD2 project. The releases that are mentioned in this section are the bigger milestone releases on Publicdata.eu and include the main releases of CKAN which have been made possible by the development on Publicdata.eu.

2.1 Early Releases

These releases were reported in Deliverable 9.1.1 and 9.1.2:

- **Alpha Launch - January 2011** - An alpha version of the site was launched in January 2011\(^8\), as part of Work Package 1 on requirements gathering, prototyping and design. Feedback from this has helped us identify improvements, and the site was subsequently moved to using the CKAN software.

- **Beta Release - June 2011** - After releasing an experimental data catalogue federation and scraping front end in January 2011, this was the first iteration of the site based on CKAN, our data management system. While the basic functionality is still that of a read-only dataset search, a lot has changed behind the scenes.

- **Upgrade - March 2012** - In March 2012 we upgraded Publicdata.eu to CKAN version 1.6 - adding the data preview functionality (powered by Recline), improvements to search, interface improvements to dataset pages, newly added resource (file) pages and group pages.

- **Upgrade CKAN 2.0 – May 2013** - The main release done in the reporting period was around the personalization features of publicdata.eu. This was part of a major overhaul of CKAN which resulted in the CKAN 2.0 release, which was announced in beta in February 2013 and after more testing and bugfixing it was officially released in May 2013.

2.2 Work done in Sept 2013 – Aug 2014

In the last year of LOD2 we released two more versions of CKAN – 2.1 and 2.2. While CKAN 2.1 was officially released in August 2013, it was deployed on publicdata.eu in September. This version added new features, including improved previews for text files, a new redesigned dashboard and significant improvements to the documentation. More details are available from the changelog.

Version 2.2 of CKAN was released in February 2014. This version included numerous bugfixes and improved documentation, to make it easier for other organisations to use and deploy CKAN. Full information is available from the changelog.

---

3. Technology

3.1 CKAN

Publicdata.eu is powered by CKAN (http://ckan.org) - an open source data portal platform developed by Open Knowledge. CKAN is a mature project with a growing community of users, and it powers both official and community-driven data catalogs around the world.

CKAN is built with Python on the backend and Javascript on the frontend, and uses The Pylons web framework and SQLAlchemy as its ORM. Its database engine is PostgreSQL and its search is powered by SOLR. It has a modular architecture that allows extensions to be developed to provide additional features such as harvesting or data upload.

CKAN uses its internal model to store metadata about the different records, and presents it on a web interface that allows users to browse and search this metadata. It also offers a powerful API that allows third-party applications and services to be built around it.

3.2 Additional software and tools

CKAN DCAT extension

We have developed a CKAN extension which provides plugins that allow CKAN to expose and consume metadata from other catalogs using documents serialized with DCAT (available via the Github9). The Data Catalog Vocabulary (DCAT) is "an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web"10. Both the extension and the serializations and protocol are a work in progress and development will continue after the LOD2 project ends.

The DCAT extension offers the following Implementation Guidelines:

- A serialization format for dataset metadata in RDF/XML and JSON format, both based on standard DCAT properties.
- A simple mechanism for exposing a catalog metadata dumps, with optional methods for pagination and filtering.
- In terms of CKAN, this extension offers harvesters that allow importing similar dumps in RDF/XML or JSON from other catalogs to create CKAN datasets (dcat_json_harvester and dcat_xml_harvester plugins)

On the roadmap are endpoints for (paginated) dumps of all the catalog's datasets metadata in RDF/XML and JSON format. (dcat_json_interface plugin). We also intend to develop individual endpoints for describing a dataset metadata in RDF/XML and JSON format. (Note: CKAN core already offers a RDF/XML representation; how these two fit together has not been decided yet).

Recline Data Explorer (http://reclinejs.com/)

In 2013, we integrated a preview functionality for certain file types like CSV files on Publicdata.eu. This is powered by Recline, a Javascript library that offers both a powerful Data Explorer and a set of extensible of data components - data grid, graphing, and data connectors. Recline continues to be developed separately from CKAN by a small community of enthusiasts at Open Knowledge Labs11.

9 https://github.com/ckan/ckanext-dcat
10 http://www.w3.org/TR/vocab-dcat/
11 http://okfnlabs.org/projects/
Custom-built datacatalogs.org

The code of the latest version of Datacatalogs.org has been built on a custom codebase using Node.js and is available on Github\(^\text{12}\). We intend to continue development on this new version after LOD2 ends in collaboration with other developers and editors who have been actively involved with maintaining the site.

4. Harvested Catalogues

The dataset information on Publicdata.eu is gathered from national data catalogues from across the EU. The majority of them are other CKAN instances, and the CKAN harvester makes it particularly easy to pull metadata from these catalogues.

In the first versions of Publicdata.eu we have scraped open data portals and websites that did not use dedicated software that published an API to the datasets. We have found this to be unsustainable, because some of these websites change rapidly which breaks the harvester and links to the datasets already stored. At the same time, we found that many new official data catalogues were being created using CKAN, and standards are emerging for other APIs to access the datasets. For example the Open Data Support project has initiated a working group to create a DCAT application profile for data portals in Europe. Our work in the last two years has therefore mostly focused on integrating catalogues which are built on CKAN, and our standards work has focused on being able to import and export metadata that adheres to the DCAT standard.

New data catalogues are constantly being created and we know there are many catalogues which we aren’t currently federating. With this in mind we have created datacatalogs.org.

4.1 Datacatalogs.org

Datacatalogs.org aims to be the most comprehensive list of open data catalogs in the world. It is curated by a group of leading open data experts from around the world - including representatives from local, regional and national governments, international organisations such as the World Bank, and numerous NGOs.

The site currently lists 123 official data catalogues from within the EU and we have worked to integrate more of these into Publicdata.eu, giving priority to the national data portals.

In the last year we migrated the site from an old CKAN instance to a modern User Interface based on a Google Spreadsheet, which allows for much more flexibility in presenting and using the data, as we felt that CKAN was not the right tool for this kind of information. We continue work with the Open Government working group and the enlisted experts to keep the information up-to-date.

\(^{12}\) https://github.com/okfn/datacatalogs.org
In order to make sure the outputs of the work on Publicdata.eu remain used and useful, we have focused on several streams of work around the sustainability of the software and specifications developed as part of LOD2. To this end, we have started work on developing a clearer governance structure around the development of the CKAN software. This will ensure the sustainability of the software developed by OKFN under LOD2 does not need to rely on OKFN after LOD2 ends.

We have conducted a deployment survey (which is still open) to gather more information about where CKAN is being used. To date, we are aware of more than 100 CKAN deployments globally, including those of many major national governments including the US, UK, Canada, Australia and Brazil. This widespread interest in CKAN provides a fertile basis to ask for sustainable contributions from other organisations to help develop CKAN thrive as an Open Source project.

As one of the first actions, we started to expand the technical team responsible for developing CKAN by including members outside of OKFN. The technical lead on the team is currently Ian Ward, who is working for the Canadian national open data portal, and two other core developers on CKAN who are not contracted by OKFN have joined the team. We believe that this development provides an excellent basis for further opportunities for Publicdata.eu. Regardless of what company will become responsible for the portal, its reliance on the open source software CKAN will not depend on the developers working with OKFN. As such, CKAN is delivering on the promise of Free and Open Source Software.

The first announcement of CKAN Association was in March 2014. The CKAN Association has its formal institutional home at the Open Knowledge Foundation but is autonomous and has its own independent governance, in the form the Steering Group which is drawn from major CKAN stakeholders. The Open Knowledge Foundation, who are the original creators of CKAN, will continue to contribute to...
CKAN at all levels but the Association allows others – from government users to suppliers of CKAN services – to have a formal role in the development of the CKAN project going forward.

The CKAN Association has “members”. Membership is a way for individuals, companies and organisations to support the CKAN Project and be recognised for doing so. By becoming a member one is helping to ensure the long-term sustainability of CKAN. Members are expected to contribute resources – either through contributing money or providing in-kind resources such as staff time. Members receive recognition for their contribution through display on the website, participation in events etc. There are several tiers of membership, which is further explained on the CKAN website.

At the point of writing in August 2014, we have confirmed an initial round of members including several governments (including UK, Canada and US) and several CKAN vendors. We have established the Steering Group (to be announced formally in early September), and held our first CKAN Association community meeting. There is a collaborative drafting process to develop a Business Plan for CKAN Association and we are committed to furthering work on making the CKAN Association thrive.

A few other pieces of work are supporting our work in sustaining the CKAN outputs of the LOD2 project:

- We are working on a CKAN extensions registry to list the popular CKAN extensions
- We moved the code for CKAN out of the OKFN organisational repository and into an independent CKAN organisation on Github

We are confident that all of these actions help with the sustainable development of CKAN and are hopeful that further work on Publicdata.eu will build on the work done in the LOD2 project.

6. Key Statistics on Publicdata.eu

Visits on publicdata.eu increased dramatically over the last year, furthering the trend of the third year. We list the statistics below:

- As of 22 Aug 2014, publicdata.eu contains 48,555 datasets (compared to 20,262 in Sept 2013)
- These datasets are harvested from 29 data catalogues (27 data catalogues in Sept 2013 – old catalogues have been removed and new ones added)
- The portal has received a total of 69,752 visits between 21 Aug, 2013 and 21 Aug, 2014, compared with 42,359 the year before (increase of 64%)\(^{13}\)
- The portal has received 60,717 unique visitors between 21 Aug, 2013 and 21 Aug, 2014 (34,600 the year before, increase of 75%)

When we look at the visits over the course of the four years, it is clear that the popularity has steadily been increasing, as demonstrated in figure 2.

\(^{13}\) This includes all visits, so also developers and visits from the wiki transformation engine.
In the figure below, the number of visits per day is shown which tells an interesting story about when people visit the site. Visits are consistently higher on weekdays, lower on the weekends, and there is a significant dip over the Christmas holiday period. This indicates that the bulk of the visitors come to the site during work time, rather than when they are off.

Figure 2 - Visits on publicdata.eu since the portal was first launched.

The figure below shows the number of visits per EU country over the last 12 months. The UK is clearly the most popular country and other bigger countries also are showing many more visitors to the site. It is interesting to note that some smaller countries are overrepresented in the visits, like the Netherlands and Slovakia.

Figure 3 - Number of visits per day
Total number of visits per EU country
(Aug 2013 - Aug 2014)
7. Data Lifting and Harmonization

The main aim of the Publicdata.eu portal is to gather metadata about European datasets in one place, enabling users to find more relevant information. However, the datasets published by public bodies inside Europe are not preserved in semantically-rich formats and thus data lifting (e.g. reuse or integration) cost is high. To lower this barrier we build CSV2RDF extension for Publicdata.eu portal, which enables users to perform smart transformation of tabular data (i.e. CSV) to semantically-rich format (i.e. RDF) as well as collaborate with other users.

The transformation of tabular data in CSV2RDF application can be divided in several steps. On the first step CKAN crawler looks for tabular datasets, saves them to the CSV2RDF application server and validates them (e.g. some of the datasets are not available due to broken links or servers’ downtime). On the second step CSV2RDF application parses validated datasets and creates default mappings for transformation. The mappings are saved to special Mappings Wiki\(^{14}\), where any user can edit them. The first two steps are executed regularly as new tabular datasets appears. The third step includes user interactions with CSV2RDF application (crowd-sourcing): improving the default mappings, creation of new mappings, data transformation and integration. The overview of the CSV2RDF application architecture is depicted in Figure 4 (introduced steps are highlighted).

To discover the data user can directly use Mapping Wiki. The Mapping Wiki is based on MediaWiki and uses the Semantic MediaWiki (SMW) extension, which enables semantic annotations and embedding of search queries over these annotations within wiki pages. Wiki pages include mappings and metadata about recent transformation. Each mapping is rendered using the RelCSV2RDF template into a human-readable description of the parameters including links for transformation rerun and RDF download. 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. For each property we created a page in the mapping wiki with the list of dataset resources that utilize the corresponding property. Metadata about recent transformation is rendered using CSV2RDFMetadata template, exposing metadata such as number of lines in tabular data file, number of triples after transformation, transformation time

\(^{14}\) http://wiki.publicdata.eu
etc. Using the metadata user can query, for example, for the datasets with the most number of columns. Also, Mappings Wiki are available publicly through the SPARQL endpoint\textsuperscript{15}. In order to navigate to the Mapping Wiki from Publicdata.eu every dataset and resource page on Publicdata.eu has an RDF link as depicted in Figure 5 and 6.

![Figure 5 - Publicdata.eu portal. Dataset view. Navigation to the Mappings Wiki.](image)

![Figure 6 - Publicdata.eu portal. Resource view. Navigation to the Mappings Wiki.](image)

User can edit mappings on the Mapping Wiki using several different interfaces, depending on the dataset type. We distinguish two dataset types, based on the structure of the dataset contents. Contents of the first type are organized in way, where each row of data can be represented as an object (or RDF entity). For example, in Figure 7 data about transactions are represented. Each row represents one transaction, for example, data in the first row will be represented as:

\begin{verbatim}
@prefix pdeu-o: <http://wiki.publicdata.eu/ontology/> .
@prefix current-resource: <http://data.publicdata.eu/de023b34-b229-4392-9aec-93cbbd424da31/> .

current-resource:1 pdeu-o:Description "Cleaning and hygiene" ;
  pdeu-o:Final%20Transaction%20Amount "583.68" ;
  pdeu-o:Merchant%20Name "BUNZL CLEANING & HYGIENE" ;
\end{verbatim}

\textsuperscript{15} http://wiki.publicdata.eu/sparql
Listing 1. Example of object-per-row data.

Conversion to RDF for this datatype 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. We created a template called RelCSV2RDF, which defines the following parameters (line numbers correspond to Figure 8):

- (line 2) name: a string, which identifies the mapping and must be unique within the scope of one resource;
- (line 3) header: an integer or an integer range, which determines the position of header row(s);
- (line 4-5) omitRows and omitCols: integer ranges, which determine rows and columns to be omitted from the conversion;
- (line 5) delimiter: a symbol, defining the column delimiter for the tabular data file;
- (lines 6-13) col1, col2, col3 etc.: strings, which specify RDF properties to be used for the conversion of each table column;
- (line 14) class: URI, defines the class for instances generated for this dataset.
Figure 8 - The Mappings Wiki mapping example with wiki markdown view.

Therefore user can edit mappings directly on the Mappings Wiki using wiki markdown. In this case user has to conform to the Mappings Wiki syntax and look up properties and classes for the mappings by herself. To support user with these tasks we developed extension to the Mappings Wiki interface, which reads the mappings from the Mappings Wiki and renders it in a GUI, where user can search for properties and classes from LODStats and other third-party services (c.f. Figure 9).
The GUI for the Mappings Wiki allows user to easily define class for the table as well as properties for each column. The suggestions for classes and properties are derived from third-party services such as LODStats or LinDA Metadata and Vocabulary Repository. After the completion of the mapping user can save it back to the Mappings Wiki or start transformation from the GUI.

In order to obtain an RDF graph from a table T and a mapping 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 too often turned out to be simply labels rather than meaningful type names. Listing 3 shows the default mapping expressed in Sparqlify-ML syntax.

```
Prefix pdd: <http://data.publicdata.eu/>
Prefix pdo: <http://wiki.publicdata.eu/ontology/>

Create View Template DefaultMapping As
Construct {
?s ?p1 ?o1 ;
?p2 ?o2 
} With
?s = uri(concat(pdd:,’csv-path/’,?rowId))
?p1 = uri(concat(pdo:, ?headingName1))
?o1 = plainLiteral(?1)
?p2 = ...
```

Listing 3. Default mapping expressed in Sparqlify-ML.

CSV files using the same column header will produce RDF containing the same properties. We argue, that in the majority of the cases this behaviour is desirable, especially, if multiple
datasets were exported to CSV from the same backend system and have the same structure and headers. However, this automatic mapping can also result in incorrect property identification in cases, where columns in CSV files have the same header label, but different meaning. Our crowd-sourcing approach enables to quickly resolve such problems once identified.

Mappings can be reused completely or partially if CSV files have similar header items. By reusing the same mapping CSV files are transformed to RDF data, which can be easily integrated. This feature is supported by Mappings Wiki GUI. For example, we transformed CSV file “2010 September Return” from Publicdata.eu dataset “Spend over £25,000 in Southampton City PCT”\(^{16}\). The transformation mapping used in this example is represented in Listing 4.

```
{{RelCSV2RDF|
  name = csv2rdf-interface-generated-with-datatype |
  col8 = Amount^^xsd:float |
  col6 = Supplier |
  col7 = Transaction Number |
  col4 = Expense Type |
  col5 = Expense Area |
  col2 = Entity |
  col3 = http://purl.org/dc/elements/1.1/date^^xsd:dateTime |
  col1 = http://data.lirmm.fr/ontologies/passim#department |
  header = 1 |
  omitCols = -1 |
  delimiter = , |
  omitRows = -1 |
  class = http://dbpedia.org/ontology/Organisation |
}}
```

Listing 4. The example mapping for transformation of “2010 September Return” CSV file.

For the example CSV file 16 other CSV files with the similar header structure were detected and converted to RDF using the same mapping. All the RDF data after conversion is loaded in a triplestore. As a result user is able to query all 17 CSV files as RDF data. For example, using the SPARQL query defined in Listing 5 and Sgvizler visualization tool user is able to visualize the amount of money spent by Southampton City PCT COM over time, while in original data each month is represented as a separate CSV file (c.f. Figure 10).

```
select distinct ?date ?amount
WHERE {
  FILTER (regex(?entity, 'Southampton City PCT COM') && ?amount > '0'^^xsd:float) }
```

Listing 5. SPARQL query for retrieving payments from Southampton City PCT COM.

---


A special subtype of object-per-row data is spatial data that is tabular data containing latitude and longitude columns. The processing of such data is performed with the same interfaces. By specifying commonly used properties for latitude and longitude the data become eligible to be visualized on the map by, for example, Facete. Facete, depicted in Figure 11, is a 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.

Figure 10 - Sgvizler visualization over the integrated RDF data for “2010 September Return”.

https://github.com/GeoKnow/Facete, Screencast: http://youtu.be/VzEvFJs89Wc
Figure 11 - Screenshot of Facete showing data about allotments in South East London.

The contents of second dataset type are organized in a way, where every cell should be represented as an object (or RDF entity). For example, in Figure 12, data about crime prosecutions are represented. Each cell represents one statistical entry, for example, cell with value 1887 will be represented as:

```reason
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix res: <http://ontowiki.vhost.tld/csvimport//353947211e62bddc1ac589f2e4f1e182/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
res:c1-r1 res:Women res:r1-c0 ;
res:YearVolume res:r0-c1 ;
res:m "1887"^^xsd:integer ;
qb:dataSet res:ds ;
a qb:Observation .
```

Listing 6. Example of object-per-cell data.

Figure 12 - Publicdata.eu portal. Data view. Example of object-per-cell data.

To convert this type of data to RDF we utilize csvimport OntoWiki plugin\(^{19}\). Once statistical data represented in a tabular form converted to the RDF DataCube vocabulary, the user is able to discover the data using CubeViz\(^{20}\), the RDF DataCube browser. CubeViz generates facets as illustrated in Figure 13 according to the RDF DataCube vocabulary:

1. Selection of a DataCube Dataset,
2. Selection of a DataCube Slice,

\(^{19}\)http://aksw.org/Projects/CSVImport.html, usage presented in screencast: http://youtu.be/lb8b8YWU2i8

\(^{20}\)http://aksw.org/Projects/CubeViz
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.

![Figure 13 - Screenshot of CubeViz with faceted data selection and chart visualization component.](image)

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 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.
8. Future of publicdata.eu

The European Commission released a tender for the deployment of the next generation of Publicdata.eu in July 2014. One of the requirements for the tender is the use of Free and Open Source Software (FOSS) and the sustainability of that software is one of the criteria for evaluation.

We hope and expect that with the ongoing development on CKAN over the last 4 years, much of which as part of the development of the prototype publicdata.eu, the software that underpins the current version of Publicdata.eu will be considered by whoever will be awarded the contract. We have worked on a better sustainability model around how the software is being developed with the CKAN Association, to ensure that whoever will win the tender will have the opportunity to engage with the development team of CKAN at OKFN and will be able to build on the sustainable software stack that the Publicdata.eu prototype has provided. Work on CKAN and DCAT will continue regardless of what will be deployed on the new version of Publicdata.eu. Because of key buy-in from major partners in the CKAN Association, the future of the software developed partly as within Publicdata.eu is looking bright!