Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube Vocabulary

Pilar Escobar[a,*], Gustavo Candela[a], Juan Trujillo[a], Manuel Marco-Such[a], Jesús Peral[a]

[a] *Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, carretera San Vicente s/n 03690 San Vicente del Raspeig, Alicante (Spain)*

Abstract

Most organisations using Open Data currently focus on data processing and analysis. However, although Open Data may be available online, these data are generally of poor quality, thus discouraging others from contributing to and reusing them. This paper describes an approach to publish statistical data from public repositories by using Semantic Web standards published by the W3C, such as RDF and SPARQL, in order to facilitate the analysis of multidimensional models. We have defined a framework based on the entire lifecycle of data publication including a novel step of Linked Open Data assessment and the use of external repositories as knowledge base for data enrichment. As a result, users are able to interact with the data generated according to the RDF Data Cube vocabulary, which makes it possible for general users to avoid the complexity of SPARQL when analysing data. The use case was applied to the Barcelona Open Data platform and revealed the benefits of the application of our approach, such as helping in the decision-making process.

Keywords: Linked Open Data, Multidimensional Modelling, Conceptual Modelling, RDF Data Cube Vocabulary, Semantic Web, Big Data

*Corresponding author
Email address: mpilar.escobar@ua.es (Pilar Escobar)

1. Introduction

The technological advances made in the last few decades have enhanced and connected the entire globe. In June 2018, more than 4 billion users worldwide were connected to the Internet, which is approximately 55% of the world's population [1]. This scenario has generated a huge volume and variety of data that are inadequate for traditional computers, and which are known as Big Data [2].

The publication of Open Data has recently attracted great interest among the research community [3]. Open Data are data that can be freely accessible and that become usable when made available in a common, machine-readable format, thus allowing them to be automatically read and processed by a computer. Tim Berners-Lee [4] proposed a new model called Linked Data to publish machine-readable information as structured data, based on RDF [5]. RDF encodes facts as triples, including a subject, a property and an object, each of which is identified by a Universal Resource Identifier (URI). Linked Open Data (LOD) are, meanwhile, Linked Data that are released under an open licence. However, querying content published as LOD effectively means understanding the semantic concepts contained in the repository and being able to write complex queries in Simple Protocol and RDF Query Language (SPARQL) [6].

Many libraries, museums, and archives are currently exploring ways in which to publish their catalogues as Open Data and to develop new interfaces that will provide the users of cultural heritage websites with a richer experience. Moreover, several cities around the world, such as Madrid, London, Barcelona, Paris or New York, have become great producers of data [7, 8, 9, 10, 11]. The publication of statistical data by means of reliable standards enables easy reuse, and the efficient management of this volume of data is crucial if the public is provided with consistent and structured information. However, all too often, Government data may be available online, but are still of poor quality, thus discouraging others from contributing to and reusing them [12, 13, 14]. In [15], seven Open Data portals are assessed obtaining a low score regarding the machine-processable indicator, which means that they publish data mostly in

PDF, DOC, XLS, and other non-processable formats. In addition, Muller et al. [16] noted that in order to exploit data sources, focused on non-conventional data, a series of challenges had to be addressed, the main one, the quality of the data.

According to the Open Data Barometer report [17], good quality Open Data needs to be: (i) Available online so as to reach the widest practical range of users and uses; (ii) Machine-readable so that large datasets can be analysed efficiently; (iii) Available in bulk so that they can be downloaded as one dataset and easily analysed by a machine; (iv) Free of charge so that anyone can access them no matter what their budget is; (v) Open-licensed so that anyone will be able to use and reuse the data.

Several openly available knowledge graphs (KGs), based on LOD concepts, have been created in parallel, such as DBpedia [18], Wikidata [19] and YAGO [20]. These KGs cover general knowledge, also known as cross-domain knowledge, rather than knowledge concerning special domains. KGs are a rich source of information to enrich datasets, since they provide structured data such as links to other authorities (e.g. GeoNames) that can be further exploited, and access to descriptions and properties in multiple languages.

The main contributions of this paper are the following: (a) the proposal of a framework to enhance the enrichment and publication of Open Data by means of multidimensional models and LOD; (b) the definition of a novel step of LOD assessment using different criteria concerning data quality; (c) the dataset exploitation providing dashboards that allow non-expert users to interact with data generated according to the RDF Data Cube vocabulary (using existing tools such as CubeViz); and (d) the evaluation of the framework by means of a case study applied to the Barcelona's official Open Data platform illustrating how the transformation process can aid in the decision-making process.

The remainder of the paper is structured as follows. Section 2 presents work related to Open Data and the publication of statistical data, while Section 3 describes our approach for use in the publishing of statistical data from disparate sources. Section 4 describes a real case scenario using the data from the

Barcelona Open Data platform, and finally, Section 5 shows our conclusions and future work.

2. Related work

In this section, we provide an overview of the concepts and previous research related to Open Data, along with the current state of Open Data reuse.

2.1. Open Data

Open Data are a key resource for social innovation and economic growth [21], and have a tremendous commercial value [22]. Providing access to data concerning public services by means of Open Data will open up a new scenario in which governments will be able to collaborate with citizens as regards, for example, the evaluation of public services. Moreover, both traditional businesses and new entrepreneurs are using Open Data not only to better understand potential markets, but also to build new data-driven products.

Many cities in the world are currently producing huge amounts of data. A discussion regarding Open Data utilisation in five smart cities (Barcelona, Chicago, Manchester, Amsterdam, and Helsinki) was presented by Ojo [23]. Dong [24] provides a detailed explanation of the datasets for each Canadian city, including the different data catalogues and their detailed characteristics. Both highlight the significance of Open Data and its resulting innovations in these cities.

Cities normally group their datasets into categories based on government activities, such as security, culture and leisure, environment, transport and city facilities [25].

The format of an open dataset refers to how the data are structured and published for humans and machines. Choosing the right format enhances their management and reuse. However, a satisfactory response to users' needs could be provided by using common formats (e.g. text files) and others that are more advanced and not so widespread [26].

In order to make data available, publishers generally organise a central catalogue in which to list the datasets. It is also possible to consider alternatives means, such as the use of an Application Programming Interface (API), which allows programmers to identify and select data from the entire data set by using a custom set of criteria, rather than downloading the entire dataset. Others initiatives propose a conceptual model which is able to originate an effective scalable e-Government ontology [27]. In addition, the interlinking of Schema.org to vocabularies has been analysed in order to enhance the enrichment process of a dataset [28].

In addition, since cities are complex systems producing huge amounts of data, there are still research challenges concerning the study of advanced techniques of visualisation and services to enable data exploration. In this context, smart city ontologies, such as Smart City Ontology (SCO), provide a powerful tool for semantics-enabled exploration of urban data [29, 30].

Free and open knowledge bases such as Wikidata[1] have, meanwhile, been growing in popularity, thus promoting the publication and reuse of Open Data. Wikidata takes an innovative approach by providing an online workflow in order to propose the creation of new properties that are discussed in a participatory manner and, if there are some supporters and a consensus is reached, the property is eventually created by an administrator.

2.2. Barriers to Open Data reuse

Today, most organisations using Open Data focus on data processing and analysis, and some of the services provided are, for instance, the transformation of raw data into actionable insights. However, much work must still be done if the full potential of reusing Open Data is exploited [31].

According to the report published by the European Commission concerning the reuse of Open Data [32], both external and internal barriers remain, which hinder re-users from standardising or automating the collection and processing

[1] https://www.wikidata.org, accessed 11-February-2019.

of Open Data. The report concludes with a series of recommendations for both the public and private sectors.

Ruijer [13] suggested that the interaction among governments, industries and universities could overcome the barriers that prevent governments from implementing new technologies and smart processes, owing to their tight budgets and human resource constraints.

In [14], data users cited that the lack of basic guidelines for the use and enrichment of the available data has a negative impact on the reuse level. They suggested the creation of a basic reuse kit including a guideline that would help them to download, connect, enrich and display released data, which could help re-users to understand how the city's open datasets could be used in a meaningful way.

Link [12] suggested factors that could reduce the impact of Open Data, such as recollection by automated tools that pose challenges as regards guaranteeing privacy, data quality, and analysing the data. When it comes to considering which dataset to use, data-quality is a crucial aspect. A number of initiatives have been undertaken in order to specify and evaluate the quality of linked data [33, 34]. The evaluation of a LOD includes several aspects such as consistency, accuracy, and completeness.

2.3. Methodologies for publishing Linked Open Data

As more Open Data are published on the Web, best practices and guidelines are also evolving. In [35], a Linked Data life-cycle workflow architecture is proposed based on four components: (1) Acquisition, (2) Ontology Learning Method, (3) RDF Store and (4) Analysis System. In [36], limitations and drawbacks of current frameworks are identified and a methodology for publishing LOD with the use of cloud computing is proposed.

The W3C Government Linked Data Working Group proposes a guide to aid in the access and re-use of Open Government Data [37]. In addition to this guide, several publications propose life cycle models, that share common activities, such as specifying, modelling and publishing data in standard open Web

formats. Hyland [38] provides a lifecycle consisting of the following activities: (1) Identify, (2) Model, (3) Name, (4) Describe, (5) Convert, (6) Publish, and (7) Maintain. In Villazón [39], the authors propose a preliminary set of methodological guidelines to assist in the generation, publication and exploitation of Linked Government Data. Their life cycle consists of the following activities: (1) Specify, (2) Model, (3) Generate, (4) Publish, and (5) Exploit. They capture the tasks that are required in a traditional information management workflow, but provide different boundaries between these tasks [37].

We can conclude this section emphasizing that we identified a lot of common features and functionalities between compared frameworks. However, some key features were omitted or not used, such as the use of external repositories as knowledge base for data enrichment (for instance, Wikidata or GeoNames), and the inclusion of a step to perform the assessment of LOD.

2.4. Linked data and multidimensional datasets

In the topic of LOD, multidimensional models are the combination of different datasets, which enable the application of evaluation techniques by means of statistics and indicators [40, 41]. According to the W3C, a statistical data set comprises a collection of observations that can be organised into a set of dimensions, attributes and measures, known as components [37].

The RDF Data Cube vocabulary [42] is a W3C recommendation for the publication of multidimensional data, such as statistics, on the Web. It specifically defines the dimensions, attributes and measures used in the dataset and builds upon existing RDF vocabularies (for example, SKOS, SCOVO, Dublin Core, FOAF, etc.). The Data Cube vocabulary is compatible with SDMX (Statistical Data and Metadata eXchange), an ISO standard used to exchange and share statistical data and metadata among organisations.

Literature also contains some examples of the use of multidimensional models and Linked Data [43, 41]. In [44, 45], a new vocabulary is proposed as an extension of RDF Data Cube vocabulary that supports advanced OLAP operations, such as rollup, slice, dice, and drill-across, using standard SPARQL

queries. Several national statistics institutes including Italy,[2] Ireland,[3] Greece,[4] Scotland,[5] UK[6] and Japan [46] provide their statistical data as LOD based on the RDF Data Cube vocabulary. In [47], the publication of official pension statistics as LOD based on the RDF Data Cube vocabulary illustrates how the data is reused in applications and how it contributes to statistical indicators in combination with other LOD. In addition, AirBase is the European air quality dataset maintained by the Environmental European Agency which represents air pollution information as an RDF data cube, which has been linked to the YAGO and DBpedia knowledge bases [48].

In parallel, new applications for the visualisation and exploration of statistical data based on the RDF Data Cube vocabulary have recently been published. IT-infrastructures often have strict requirements regarding the integration of new applications. Traditional client-server applications depend on the availability of the server-side part. In contrast, CubeViz.js [49] is a client-side only application that allows connections to be made to a SPARQL endpoint or a file dataset. CubeViz.js is based on the RDF Data Cube vocabulary and is able to process the Data Cubes provided by a self-maintained SPARQL endpoint, along with Data Cubes that are published as Turtle or JSON files. Moreover, [50] propose four methods for linked data viewing identifying potential uses cases of a dataset such as, an overview of queries and different tools to allow data to be visualized.

However, some challenges remain related to the creation of cubes as linked data and approaches to addressing them, highlighting the difficulties to integrate different sources and the development of generic software tools [51].

[2] http://datiopen.istat.it/index.php?language=eng
[3] http://data.cso.ie/sparql
[4] http://linked-statistics.gr/sparql
[5] https://statistics.gov.scot/
[6] http://statistics.data.gov.uk/sparql

2.5. Findings and contributions of our proposal

After reviewing the previous work, we identified a lot of common features between the frameworks oriented towards publishing and exploiting linked data. However, some key features were omitted or not used. We present below the main challenges and open issues in this area:

- The inclusion of a step to perform the assessment of LOD.
- The use of external repositories as knowledge base for data enrichment.
- The improvement of data exploitation and visualization.
- The analysis of the different data sources to automate their integration.

Below, we summarize the main contributions presented in this paper:

- The proposal of a generic framework to enhance the enrichment and publication of Open Data by means of multidimensional models and LOD.
- The definition of a novel step of LOD assessment using different criteria concerning data quality.
- The enrichment of the original dataset by using links to external repositories (such as Wikidata and GeoNames).
- The dataset exploitation providing: (a) dashboards that allow non-expert users to interact with data generated according to the RDF Data Cube vocabulary (using existing tools such as CubeViz), and (b) a public SPARQL endpoint for expert users.
- The evaluation of the framework by means of a case study applied to the Barcelona's official Open Data platform illustrating how the transformation process can aid in the decision-making process.

3. The framework for publishing Linked Open Data

In the following subsections, we describe each step of our framework based on the life cycle of Villazón [39] which includes the main methodological guidelines oriented towards publishing and exploiting linked data. Our approach enhances the original process of Villazón by including an additional step of LOD assessment based on the methodology proposed by [34] and adapted to the specificities of data cube repositories. Furthermore, the enrichment of the original dataset by using connections to external repositories has been carried out. In addition, to facilitate the repository exploitation, dashboards and a public SPARQL endpoint have been made available. In Figure 1 the proposed framework is shown.

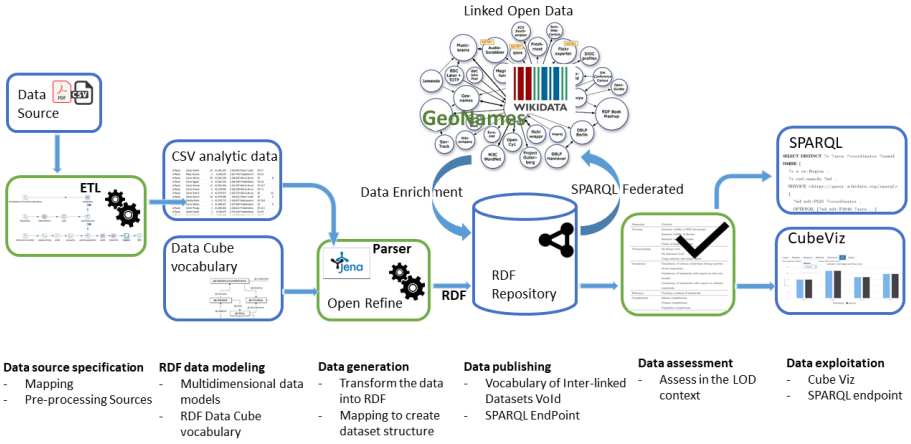

Figure 1: The framework for publishing Linked Open Data.

3.1. Data source specification

The format of a dataset refers to how data are structured and published for humans and machines. Choosing the right format enhances management and reuse. While the most common format used by organisations to publish data is CSV (Comma Separated Files), which is simple to understand, highly reusable

and machine-readable, more advanced approaches use XML, RDF and JSON, thus providing a higher level of information in terms of semantics [26]. However, in some cases, statistics are more understandable and readable when using XLS as a format, but its macros and formulas may be hard to handle.

In addition, it is not possible to guarantee the homogeneity of Open Data data across institutions owing to the variety of data formats, vocabularies and external repositories. Common problems appear, such as textual errors, typos, abbreviations, languages, a lack of information and the disambiguation of locations [52]. The pre-processing step, therefore, generally includes a set of parsers (e.g. implemented in Java, Python or using Extraction, Transform and Load tools [53]) in order to normalise the information contained in the source data.

Our approach is based on the development of Extraction, Transform and Load (ETL) processes designed by means of Pentaho Data Integration (Kettle),[7] which is a modern data integration platform that allows access to and the preparation, combination and analysis of unstructured data, in order to normalise the data obtained from heterogeneous data sources.

It is important to note that although data sources may differ across institutions, our approach is generic in order to facilitate its application to any domain. This process requires the identification of commons points in the data sources in order to join them. Once the original data sources are treated as a whole, several additional tasks are required such as cleaning and normalising the data. As a result of this semi-automatic process, a unique file with the integrated information is returned which is finally used to create the RDF.

3.2. RDF data modelling

This step covers the transformation from the original sources into the form of a multidimensional data model, including components such as dimensions, measures and attributes. The dimension components are used to provide infor-

[7] https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial, accessed 11-February-2019.

mation about the observations, i.e., the time to which the observation applies, or a geographic location at which the observation occurs. The measure components represent the fact being observed, such as the population of a region or per capita income. The attribute components serve to qualify the observations, i.e., the specification of the units of measures, along with, additional metadata, such as the status of the observation (e.g. hidden, checked).

Multidimensional data models, may, in particular, have different relational representations, including the star schema and the snowflake schema [54], both of which use dimension tables to describe data aggregated in a fact table.

The most common is the star schema, whose main feature is that its dimension tables are not normalised. This representation looks like a star, in which the dimension tables surround the central fact table. The granularity inside each dimension is also determined by the need for details. With regard to the snowflake schema, the most important difference is that the dimension tables are normalised and the hierarchies are divided into separate tables, thus allowing a better understanding of the classification levels defined in the dimension.

However, in order to allow the linking of this data to external knowledge bases, it is necessary to transform the data into RDF. In the case of the ontologies, the W3C Government Linked Data Working Group recommends, as far as possible, the reuse of standardised vocabularies to facilitate the inclusion and expansion of the Web of data [37].

The RDF Data Cube vocabulary uses the class *qb:dataSet* to identify a collection of observations typed as *qb:Observation*. Dimensions, attributes and measures are represented as RDF properties, typed as the abstract *qb:ComponentProperty* class, which in turn has sub-classes *qb:DimensionProperty*, *qb:AttributeProperty* and *qb:MeasureProperty*. The dimension components serve to identify the observations, such as the time at which the observation occurs or the geographic location. The measure components represent the fact being observed, while the attribute components enable the qualification and interpretation of the observed values by adding metadata concerning units of measure or the status of the observation.

Publishing multidimensional data by means of the RDF Data Cube vocabulary has several important benefits. Rather than providing consumers with static files such as CSV and PDF, the dataset is published as a machine readable and non-proprietary format. Moreover, individual observations are addressable, thus allowing third-party usage by creating references. Moreover, many components and tools are built upon the RDF Data Cube vocabulary, which enables its adoption and simplifies the set-up process and adjustment.

3.3. Data generation

This step includes the transformation of the source data into a machine readable language, i.e., RDF, thereby providing interoperability and links to other datasets. The transformation may be carried out in a batch or in a graphically-aided manner.

For example, Jena[8] is a Java API that can be used to create and manipulate RDF graphs, and provides classes to represent graphs, resources, properties and literals.

Moreover, OpenRefine[9] is a standalone open source desktop application for data cleanup and transformation to other formats. OpenRefine allows us to automatically transform raw data into a machine readable language, thus enabling graphical mapping from a project onto an RDF skeleton and its subsequent exportation in RDF format. The RDF schema alignment skeleton specifies how the RDF data will be generated from the source data. The cells in each record of the data will be placed in nodes within the skeleton.

Datasets become more useful and reusable when they are closely interlinked with other collections. These links are described by means of the *owl:sameAs* relationship and they contribute to the rich connectivity promoted by LOD. The interlinking process normally takes place in two steps: (i) an automatic procedure extracts the information from the data source, parses textual information

[8] https://jena.apache.org/documentation/rdf/index.html, accessed 11-February-2019.
[9] https://github.com/OpenRefine/OpenRefine, accessed 11-February-2019.

and finds the candidate links to external resources and (ii) a further manual refinement is carried out by data curators in order to validate the external links.

However, in some cases the automatic procedure can be particularly difficult and data curators are assisted by tools. For instance, the *Mix'n'match*[10] tool permits users to match Wikidata entries with a list of topics from external repositories in a fast and simple manner.

Many repositories can currently be used to enrich a dataset depending on the context. More and more systems rely on gazetteers in order to link natural language texts to geographical locations, with GeoNames being arguably the most commonly used gazetteer at present [55]. With regard to knowledge graphs, DBpedia and, more recently, Wikidata have become very popular within the community. In general, they provide an API in order to consume the data, which can be easily adopted.

Our approach is based on OpenRefine, since is a powerful tool as regards working with heterogeneous data, transforming it to a uniform vocabulary and enriching it with external repositories.

3.4. Data publishing

The rapid development of the Semantic Web has promoted an increase in RDF data on the Web. As a result, a set of techniques to store RDF data have been proposed. The efficient storage of RDF data has already been discussed in literature [56, 57]. There are several ways in which to store RDF data (commonly known as triple stores) that support data storage mechanisms, inference, update options, scalability, SPARQL endpoint and distribution, among others. Many options based on Javascript have recently been proposed [58]. However, other approaches directly use the final dataset, thus avoiding complex technical requirements such as installation and configuration.

In addition, the use of terms and properties from vocabularies to describe RDF datasets facilitates their discovery. For example, the Vocabulary of Inter-

[10] https://tools.wmflabs.org/mix-n-match/, accessed 11-February-2019.

linked Datasets (VoID) [59] is concerned with metadata related to RDF datasets.

By providing licensing information, users are aware of the conditions and terms of use. In general, this information is specified in RDF by means of relations such as `dcterms:licence` and `dcterms:rights`, either in the dataset or in a separate VoID file.

Since directly publishing the final dataset reduces complex maintenance tasks, our approach proposes the publication of RDF as a file that can be accessed by third-parties, including metadata, such as licensing information, and described by means of VoID.

3.5. LOD assessment

Färber et al. proposed a list of data-quality criteria to evaluate Knowledge Graphs (KGs) in the LOD context [34]. This approach employs the concepts of criteria, dimensions and categories originally proposed by previous research concerning data-quality [60].

A data-quality criterion is a function with values in the range 0–1 which scores a particular feature –such as availability and timeliness frequency. A data-quality dimension comprises one or more criteria which are grouped into categories as is shown in Table 1.

Table 1: The data-quality dimensions proposed by Färber et al. grouped by category.

Category	Dimensions
Intrinsic category	Accuracy, Trustworthiness, Consistency
Contextual category	Relevancy, Completeness, Timeliness
Representational data-quality	Ease of understanding, Interoperability
Accessibility category	Accessibility, License, Interlinking

The procedures proposed by Färber et al. [34] to evaluate every criterion have been here adapted to the specificities of data cube repositories. Section 4.5 details how each criterion has been applied to the use case.

3.6. Data exploitation

This step covers the exploitation of the dataset as a result of the transformation process. In order to exploit it to its full potential, it is necessary to provide dashboards that enable users with limited knowledge and a lack of Information Technology (IT) skills to interact with the dataset. CubeViz.js generates a faceted browsing widget that can be used to interactively filter observations that are to be visualised in charts. In addition, a public SPARQL endpoint could be enabled in order to facilitate the access and reuse the dataset.

4. A real case scenario

According to the State of European Cities Report[11] and the priorities for EU regional and urban development,[12] EU cities are on the front line as regards climate action, boosting innovation and reducing our impact on the planet.

Much work had been done to raise awareness among the public in general in order to improve the quality of life of many citizens. By exploiting and reusing the data provided by Open Data platforms, it is possible to foresee problems that may occur in the future. Data can be enriched by means of different repositories, along with being displayed using dashboards that permit decision makers to analyse how critical parts of their organisation are performing. However, an Open Data platform may often have this information as textual content but not in a structured model, thus making the search process cumbersome.

Our approach has been evaluated by using the data from the Barcelona Open Data platform in the context of Open Government Data. This case study is focused on the state of the critical cleaning spots in the city of Barcelona.[13] A critical cleaning spot is a geographical location in which various problems can be

[11] https://ec.europa.eu/regional_policy/en/policy/themes/urban-development/cities-report, accessed 11-February-2019.

[12] https://ec.europa.eu/regional_policy/en/policy/how/priorities, accessed 11-February-2019.

[13] http://opendata-ajuntament.barcelona.cat/data/en/dataset/punts-critics-neteja-barcelona, accessed 11-February-2019.

identified, such as full rubbish bins and/or those in a poor state, heavy furniture, old objects, etc. The data are obtained from a communication campaign to improve cleaning in the city of Barcelona, which was carried out in February 2017.

The details of each step of the publication process are described below.

4.1. Data source specification

This section presents the specification of the data sources according to the guidelines. As a result of this step, a CSV file format is obtained which can be automatically processed. This is a regular text file used for storage of tabular data in which the fields are separated using in this case, a comma. This is a semi-automatic step that requires a previous analysis and the identification of common points that allow the integration of data sources. The CSV file is used in the next step *RDF Data Modelling*.

In the case of the government data sources, we followed two paths:

- We reused data already opened up and published by the Barcelona Open Data platform.[14]
- We identified datasets that share common joint points (i.e. district, geographical location, etc.), and thus allowing further analysis.

Table 2 depicts the datasets that we have chosen for our case study, together with the format in which they are available, which are the input data sources in the ETL process. After an analysis of the data available on the Open Data BCN platform, we selected *urban environment* and *administrative boundary* files which are available as CSV format. We have extracted data concerning the critical cleaning spots in the city of Barcelona from the urban environment file. These include the geographical location, the neighbourhood, the visits generated by the critical point, and the reason why it is a critical point. This data has

[14] http://opendata-ajuntament.barcelona.cat, accessed 11-February-2019.

Table 2: Government datasets used in the transformation process.

Data	Provenance	Format	
Urban environment	Open Data BCN	Spreadsheet	CSV
Administrative boundaries	Open Data BCN	Spreadsheet	CSV
Distribution of income in Barcelona	Open Data BCN	text	PDF
Population	Open Data BCN	text	PDF

additionally been combined with the *administrative boundary* file in order to validate information regarding the neighbourhoods and areas of Barcelona.

In order to analyze whether *population* and *income* influenced the critical points, we combined this information with *urban environment* and *administrative boundary* files. Territorial income distribution and population data in the city of Barcelona have been extracted from a PDF file, considered as poor quality since they are not suitable to be automatically processed by a computer.

Once the source files are prepared, they can be read and processed sequentially to get data about the critical spots. Several additional tasks are required such as cleaning and normalising the data since the data in different sources files are not consistent with each other, for example: the same data may use different field names; the same field contains information of various attributes, so it is necessary to process the text to extract the data separately (e.g. the *3. la Barceloneta* text value contains the code and the name of the neighborhood).

Figure 2 shows a graphical representation of the transformation process which has three entry points that correspond to three heterogeneous data sources in terms of format and content. Finally, the data sources are combined in a single output file that will be later used to generate the RDF.

4.2. RDF data modelling

In Figure 3, we present our approach as a snowflake schema, including the fact table, which stores aggregated data (critical cleaning spots, number of visits, income per capita, population and state) created from the datasets. Surrounding

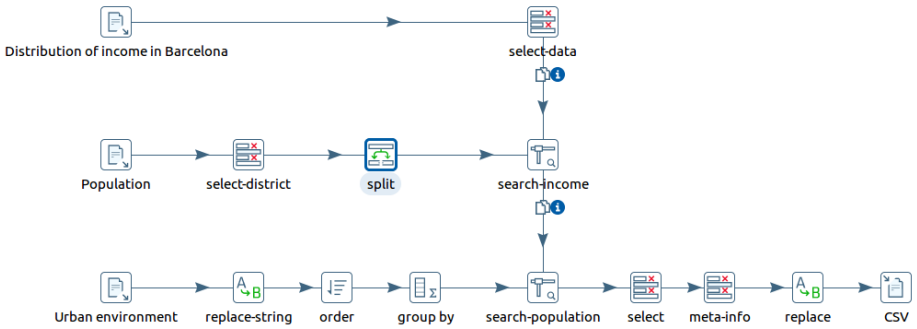

Figure 2: Transformation process based on Pentaho Data Integration (Kettle).

the fact table are the dimensions, in particular, time and region. Figure 3 is a conceptual representation of the structure that follows the final dataset for a better understanding of the use case.

The RDF data cube model obtained as a result of this step is based on the RDF Data Cube vocabulary, in which each resource is identified by an URI in order to benefit from the value of LOD. The prefixes listed in Table 3 indicate the namespaces used in the dataset. Following the design issues for the publication of LOD [4], our approach is characterised by the following structure:

- the dataset is identified by {base_URI}/dataset. A resource representing the entire dataset is created and typed as *qb:DataSet* and is then linked to the corresponding data structure definition via the *qb:structure* property.
- the data structure definition of the dataset, which includes the components such as dimensions, attributes and measures, is identified by {base_URI}/dsd and typed as *qb:DataStructureDefinition*.
- the Data Cube vocabulary represents the dimensions, attributes and measures as RDF properties. Each is an instance of the abstract *qb:ComponentProperty* class, which in turn has the sub-classes *qb:DimensionProperty*, *qb:AttributeProperty* and *qb:MeasureProperty*. For instance, time and region are typed as *qb:DimensionProperty*, while population, number of visits and critical cleaning spots are typed as *qb:MeasureProperty*.

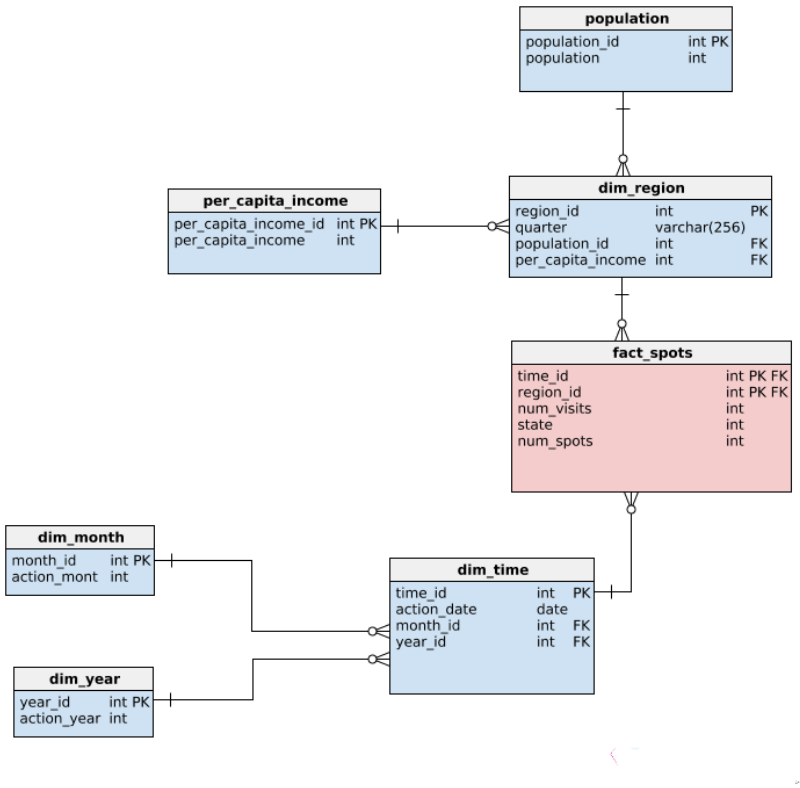

Figure 3: Conceptual representation as a snowflake schema for the critical cleaning spot in which a fact describes a number of visits at a given space/time represented as the dimensions region and time.

Table 3: Prefixes for namespaces used in the dataset.

Prefix	URI
dc	http://purl.org/dc/elements/1.1/
dcterms	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
gn	http://www.geonames.org/ontology#
owl	http://www.w3.org/2002/07/owl#
qb	http://purl.org/linked-data/cube#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
sdmx-meas	http://purl.org/linked-data/sdmx/2009/measure#
sdmx-attr	http://purl.org/linked-data/sdmx/2009/attribute#
sdmx-concept	http://purl.org/linked-data/sdmx/2009/concept#
void	http://www.w3.org/TR/void#
wd	http://www.wikidata.org/entity/
wdt	http://www.wikidata.org/prop/direct/

- each quarter is identified by the URI {base_URI}/quarter{quarter_identifier}. For instance, the quarter *Sagrada Familia*, which in the original dataset has the identifier 6, is identified by the URI *base_URI/quarter6*.

- years and months used across multiple datasets are identified by the URI {base_URI}/Yyear and {base_URI}/YyearMmonth, respectively. For instance, the year 2017 is defined as Y2017, while Y2017M1 corresponds to January 2017.

- finally, each observation is typed as *qb:Observation* and identified by a URI which contains the date followed by an auto increment number. For example, {base_URI}/201702/obs1 and {base_URI}/201705/obs2.

4.3. Data generation

This case study is based on OpenRefine to transform the source data into the RDF Data Cube vocabulary. The mappings are used to create the dataset structure, along with the observations and the components, using the appropriate URI for each element.

First of all, the resource which identifies our dataset `ex:dataset` is typed as `qb:DataSet` and additional details such as a brief description and the licence are provided. Then a `qb:DataStructureDefinition` resource is defined which references a set of `qb:ComponentSpecification` resources. Each `qb:ComponentSpecification` references a dimension or a measure by means of the property `qb:dimension` or `qb:measure`, respectively. Dimensions are typed as `qb:DimensionProperty` (e.g. `ex:geo`) while measures are typed as `qb:MeasureProperty` (e.g. `ex:perCapitaIncome`). Then, years and quarters, providing links to Wikidata and GeoNames, are defined. Finally, measures and dimensions are used to describe the observations which are typed as `qb:Observation`. Figure 4 shows an example of the mapping employed to produce the RDF dataset.

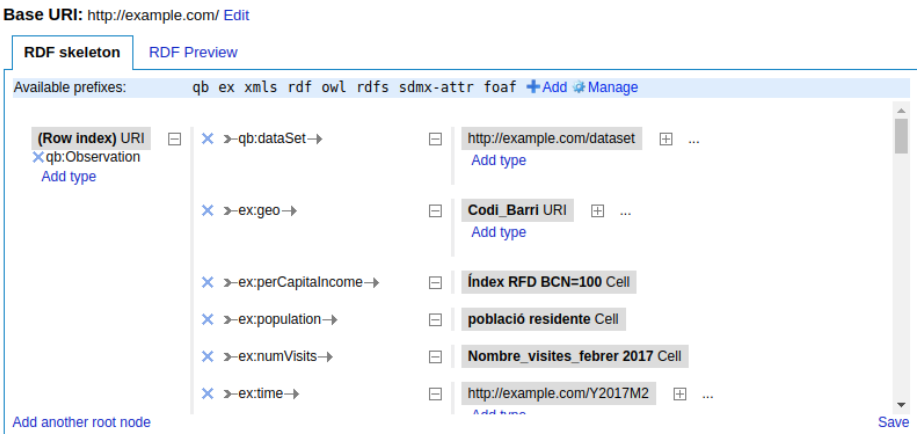

Figure 4: RDF Schema alignment with OpenRefine tool.

In order to promote data reuse and interoperability, the quarters have been manually linked to Wikidata and GeoNames, by means of an *owl:sameAs* prop-

erty. The data is eventually exported to a readable RDF syntax, such as N3, as shown in Listing 1.

```
ex:quarter1 a ex:Region ;
  owl:sameAs wd:Q1758503 ;
  rdfs:label "El Raval"@en .
ex:Y2017M2 a ex:Time ;
  skos:prefLabel "2017/february"@en ;
  skos:broader ex:Y2017Q1 ;
  skos:notation "Y2017M2" .
<http://example.com/201702/obs0>
  a qb:Observation ;
  qb:dataSet ex:dataset ;
  ex:geo ex:quarter1 ;
  ex:perCapitaIncome "74.6"^^xmls:double ;
  ex:population "47274"^^xmls:double ;
  ex:numVisits "1284"^^xmls:double ;
  ex:time ex:Y2017M2 ;
  sdmx-attr:unitMeasure ex:unit ;
  ex:state "7.3"^^xmls:double ;
  ex:criticalCleaningSpots "50"^^xmls:double .
```

Listing 1: Example of the statistics generated in N3 in which the resource `ex:quarter1` which is typed as `ex:Region` represents the quarter *El Raval*. The resource `ex:Y2017M2` represents February 2017. The last item obs0 is typed as `qb:Observation` and includes properties which use the resources defined above such as Region and Time.

4.4. Data publishing

This step includes the publication of the dataset following the LOD principles. Our approach reuses the original Creative Commons Attribution 4.0[15] licence for the Government data sources. It is stored on an RDF4J server [16] which has enabled a public SPARQL endpoint.

[15] https://creativecommons.org/licenses/by/4.0/, accessed 11-February-2019.
[16] http://rdf4j.org/

The dataset was described by means of VoID vocabulary, which helps data producers to publish metadata in a human and machine-readable format. In addition, DataHub was used as a platform on which to publish the dataset.[17]

4.5. LOD assessment

The procedures proposed by Färber et al. [34] to evaluate every criterion have been adapted to the specificities of data cube repositories, as will be explained in sections 4.5.1–4.5.11. The results obtained for the different criteria are shown in Table 4. Next, we will explain in detail the criteria related to our proposal.

4.5.1. Accuracy

- *Syntactic validity of RDF documents.* To obtain the value of this criterion, the RDF documents were validated with the RDF NTriples/Turtle Validator,[18] which confirmed that all were syntactically valid RDF documents.

- *Syntactic validity of literals.* This allows to obtain if the literals values loaded in the dataset are syntactically valid. With the same RDF NTriples/Turtle Validator it has been possible to check datatype errors. In addition, properties such as time, latitude, and longitude associated with critical cleaning spots have been checked through regular expressions.

- *Semantic validity of triples.* This criterion evaluates whether the meanings of the data in the triples are semantically correct. For example, checking if it is also available from a reliable source such as the Open Data Barcelona platform, which is an official source, or Wikidata. The value obtained indicates that a high percentage of the data is correct. For instance, the

[17] https://datahub.io/smartdataua/rdfdatacube-critical-cleaning-spots-bcn, accessed 11-February-2019.

[18] http://ttl.summerofcode.be/

Table 4: Summary of the data-quality results.

Dimension	Criterion	Value
Accuracy	Syntactic validity of RDF documents	1
	Syntactic validity of literals	0.9976
	Semantic validity of triples	1
Trustworthiness	On dataset	0.25
	On statement level	0
	Using unknown and empty values	0
Consistency	Consistency of schema restrictions during insertion of new statements	0
	Consistency of statements with respect to class constraints	1
	Consistency of statements with respect to relations constraints	1
Relevancy	Creating a ranking of statements	0
Completeness	Schema completeness	0.8
	Column completeness	0.8
	Population completeness	0.625
Timeliness	Frequency	0.25
	Specification of the validity period of statements	0
	Specification of the modification date of statements	0
Ease of understanding	Description of resources	0.12
	Labels in multiple languages	0
	Understandable RDF serialization	1
	Self-describing URIs	1
Interoperability	Avoiding blank nodes and RDF reification	1
	Provisioning of several serialization formats	1
	Using external vocabulary	0.57
	Interoperability of proprietary vocabulary	1
Accessibility	Dereferencing possibility of resources	1
	Availability of the dataset	1
	Availability of a public SPARQL endpoint	1
	Provisioning of an RDF export	1
	Support of content negotiation	0.5
	Linking HTML sites to RDF serializations	0
	Provisioning of metadata	1
Licensing	Provisioning machine-readable licensing information	1
Interlinking	Interlinking via owl:sameAs	0.12
	Validity of external URIs	1

RDF in Listing 2 shows how regions are linked to Wikidata by means of the property `owl:sameAs`.[19]

```
ex:dataset a qb:DataSet ;
  rdfs:label "Dataset"^^xmls:string ;
  rdfs:comment "Overview description about the dataset" ;
  qb:structure ex:dsd ;
  dc:publisher "Publisher Office"^^xmls:string .
ex:quarter1 a ex:Region ;
  owl:sameAs wd:Q1758503 ;
  rdfs:label "El Raval"@en .
ex:quarter2 a ex:Region ;
  owl:sameAs wd:Q17154 ;
  rdfs:label "Gothic Quarter"@en .
```

Listing 2: RDF code in which entities typed as `ex:Region` are linked to Wikidata by means of the property `owl:sameAs`.

4.5.2. Trustworthiness

- *Trustworthiness on dataset level.* The dataset is published by means of an automatic conversion to LOD. The score 0.25 is defined in [34] as data extracted from structured data sources.

- *Trustworthiness on statement level.* Vocabularies have not been included to describe the origin of the data, therefore, this criterion value is 0.

- *Using unknown and empty values.* No identifiers have been used to capture the unknown and empty values, thus the value here is 0.

4.5.3. Consistency

- *Consistency of schema restrictions during insertion of new statements.* The score obtained is 0 since the user interface does not perform checks restrictions during insertion of new statements.

[19] *El Raval* is linked to `https://www.wikidata.org/wiki/Q1758503`.

- *Consistency of statements with respect to class constraints.* The `owl:disjointWith` property has been used in order to check the class constraints. The constraints class have been checked through the SPARQL endpoint. The Listing 3 shows a SPARQL statement that checks that an entity can not be typed as *Dimension* and *Measure* simultaneously. No inconsistencies have been identified.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?e
WHERE { ?e rdf:type qb:dimension .
        ?e rdf:type qb:measure }
```

Listing 3: SPARQL query retrieving resources typed simultaneously as Dimension and Measure.

- *Consistency of statements with respect to relation constraints.* This criterion evaluates the degree of consistency of the instance with the relationship restrictions. For example, the relation `rdfs:range` indicates the type of entities that can occur in the third position in a triple. This restriction can be verified with the SPARQL query shown in Listing 4.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
SELECT distinct ?rangeType
WHERE { ?x qb:measure ?o .
?o a ?rangeType }
```

Listing 4: SPARQL query to assess the type of entities that can occur in the third position in a triple.

4.5.4. Relevancy

- *Creating a ranking of statements.* The dataset does not support the ranking of statements.

4.5.5. Completeness

In order to evaluate this dimension, it is also necessary to define a set of classes and properties listed in Table 5, which is based on RDF Data Cube[20] and DBpedia[21] vocabulary.

Table 5: Gold standard which includes the classes and properties used to evaluate the completeness criteria.

Class	Properties
Region	name
District	name
Point	name, latitude, longitude
Observation	point, per capita income, population, time, number of visits, state, unit of measure

- *Schema completeness.* It has been calculated as the ratio of the number of classes and attributes of the gold standard that exist in the dataset. A high score is obtained for this criterion because its main vocabulary is based on the RDF Data Cube vocabulary. Although the latitude and longitude properties are not included in our final dataset, they could be automatically retrieved from Wikidata thanks to the `owl:sameAs` links.

- *Column completeness.* This criterion is defined as the rate of instances which have a specific property defined, averaged for all properties in the gold standard. A value of 0.8 has been obtained since our dataset does not store information about the geographical points.

- *Population completeness.* This criterion is defined as the extent to which our dataset covers the basic population. In order to select the most popular entities per class, we have used quantitative statements. The score

[20] https://www.w3.org/TR/vocab-data-cube/
[21] http://dbpedia.org/ontology/

obtained by our dataset is low because it lacks well-known entities of our gold standard, although all of them are represented in Wikidata.

4.5.6. Timeliness

- *Timeliness frequency.* The frequency of updates was consulted by means of properties such as `dcterms:created`, and the inspection of VoID files. The score 0.25 in Färber's methodology corresponds to discrete non-periodic updates.

- *Specification of the validity period of statements.* The dataset does not use properties —such as Wikidata *end time* (`P582`)– to specify validity.

- *Specification of the modification date of statements.* No information concerning modification dates such as `dcterms:modified` or `schema:dateModified` were found.

4.5.7. Ease of understanding

- *Description of resources.* The rate of entities described with the property `rdfs:label` has been computed and found to be low.

- *Labels in multiple languages.* The string value of a property can be encoded in multiple languages by adding attributes such as `@es`, `@en`, etc. The dataset declares the language of the `rdfs:label` and `rdfs:comment` properties, in which references only to English were found.

- *Understandable RDF serialization.* Alternative encodings that are more understandable for humans than RDF include N-Triples, N3 and Turtle [61]. The dataset provides only Turtle serialization and additional formats can be obtained through the use of the SPARQL endpoint.[22]

- *Self-describing URIs.* Self-descriptive URIs contain a readable description of the entity rather than identifiers. The dataset contains a readable description of the entity class and an identifier of the resource.

[22] http://docs.rdf4j.org/rest-api/#_content_types

4.5.8. Interoperability

- *Avoiding blank nodes and RDF reification.* Blank nodes were checked by means of the `isBlank` SPARQL operator. The RDF reification vocabulary[23] is not used in the dataset.

- *Provisioning of several serialization formats.* By configuring the request, the RDF server provides results in RDF/XML, JSON-LD and Turtle which corresponds to score 1 in Färber et al. specifications.

- *Using external vocabulary.* This score was obtained by obtaining the rate of triples using an external vocabulary which correspond to 28 properties from 8 external vocabularies.

- *Interoperability of proprietary vocabulary.* This criterion determines the rate of classes and properties with at least one equivalence link to classes and properties in external vocabularies by means of properties such as `owl:sameAs`, `owl:equivalentClass`, `rdfs:subPropertyOf` or `rdfs:subClassOf`. All the classes and properties are taken from external vocabularies based mainly on qb, DC, RDF and SKOS.

4.5.9. Accessibility

- *Dereferencing possibility of resources.* A random choice of 100 URIs was requested for all the libraries by using the `application/rdf+xml` field in their HTTP header, and they all returned a correct RDF document.

- *Availability of the dataset.* The SPARQL endpoint was monitored for a period of 15 days with a 5-minute check interval. No interruptions to the service were identified.

- *Availability of a public SPARQL endpoint.* The dataset is stored in an RDF4J[24] server and the SPARQL endpoint is located at `http://data.cervantesvirtual.com/rdf4j-server/repositories/rdfdatacube`.

[23] `https://www.w3.org/TR/rdf-schema/#ch_reificationvocab`
[24] `http://rdf4j.org/`

- *Provisioning of an RDF export.* The dataset is available as a data dump based on N-Triples.

- *Support of content negotiation.* The consistency between the RDF serialization format requested (RDF/XML, N3, Turtle, and N-Triples) and that which was returned was checked. As a result, Turtle was not supported.

- *Linking HTML sites to RDF serializations.* This criterion has not been considered in our dataset since there is not an HTML website in which items are browsed.

- *Provisioning of repository metadata.* The repository can be described using Vocabulary of Interlinked Datasets (VoID) [59]. The dataset includes a VoID file with the title, description, creators and vocabularies used.

4.5.10. License

- *Provisioning machine-readable licensing information.* A license can be specified by means of the relations `dcterms:licence` and `dcterms:rights` included in either the dataset itself or a separate VoID file. Data are distributed under a Creative Commons[25] license specified by means of a `dcterms:rights` property.

4.5.11. Interlinking

- *Interlinking via owl:sameAs.* This score is obtained as the rate of instances having at least one `owl:sameAs` triple pointing to an external resource.

- *Validity of external URIs.* The number of timeouts and HTTP errors were computed when accessing a random sample of 100 URIs defined with the `owl:sameAs` relation.

4.6. Data exploitation

In order to exploit the full potential of the dataset, and avoid technical requirements such as the use of SPARQL, our approach uses CubeViz.js to

[25] https://creativecommons.org/licenses/by/4.0/

provide statistical data exploration and visualisation. The application provides a comprehensive set of features to select and visualise observations, thus assisting in the decision-making process.[26]

Figure 5 shows an example of how to enable users to interact with the data generated according to the RDF Data Cube vocabulary by means of a dashboard, which makes it possible for general users to avoid the complexity of SPARQL when analysing data. The upper menu bar allows measures, such as Population or State and dimensions, which are represented as Region (zones) and Time (months) in Figure 5, to be filtered.

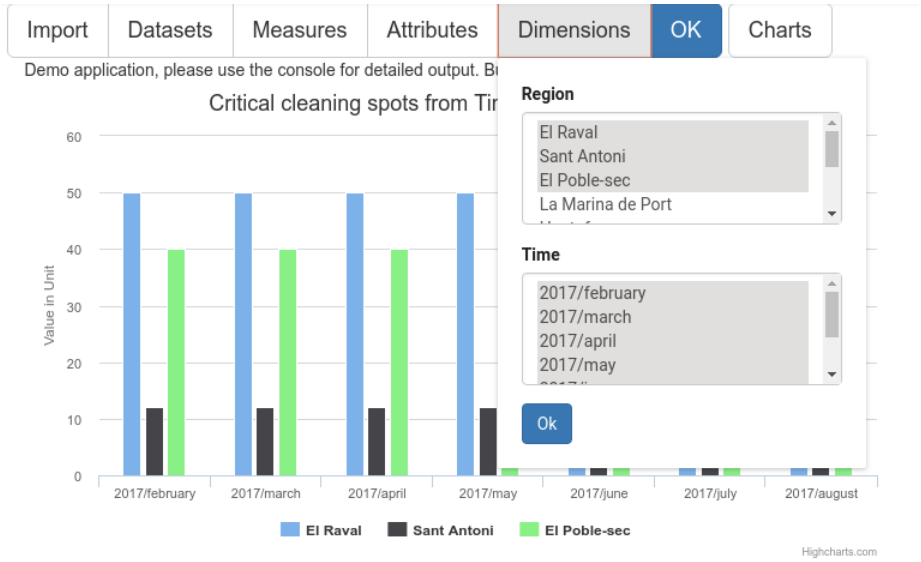

Figure 5: Critical cleaning spots in the *Region* (zones) and *Time* (months of 2017) dimensions.

Figures 6 and 7 show a graph representing critical cleaning spots and population in the *Region* (for a selection of neighborhoods) and *Time* (February 2017) dimensions. In our case scenario, we concluded that there is no correlation between the number of residents and the number of critical cleaning spots. We also concluded that the number of critical cleaning spots does not change

[26] https://smartdataua.github.io/rdfdatacube/ accessed 11-February-2019.

significantly over the year. However, we can confirm that there is a relationship between the number of critical cleaning spots and the low per capita income in areas such as *El Raval* and *El Poble Sec* (see Figure 8).

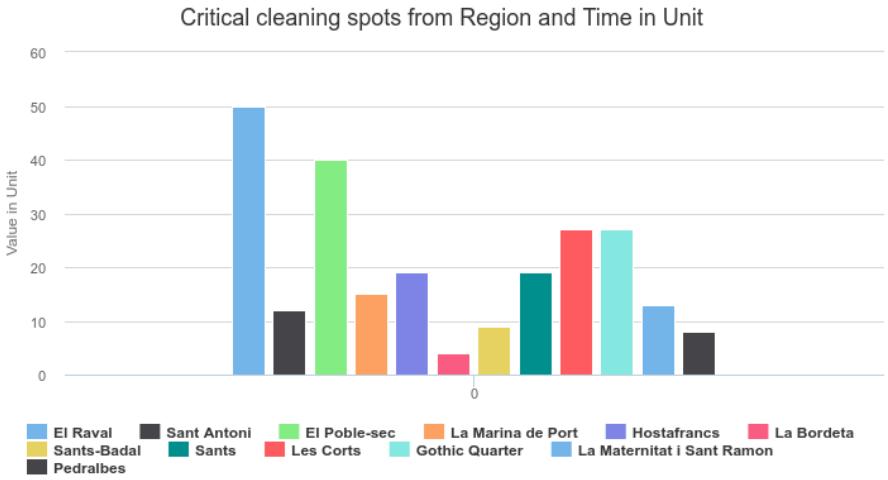

Figure 6: Critical cleaning spots in the *Region* (selected zones) and *Time* (2017/February) dimensions.

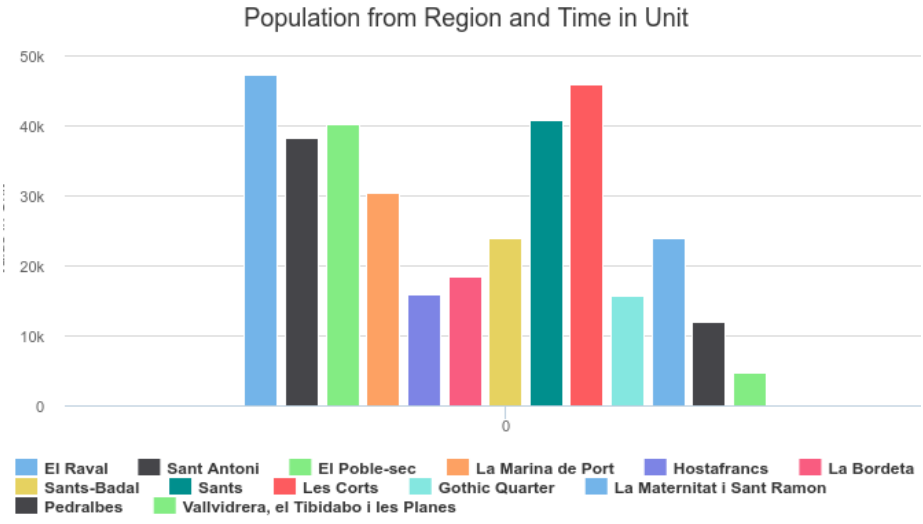

Figure 7: Population in the *Region* (selected zones) and *Time* (2017/February) dimensions.

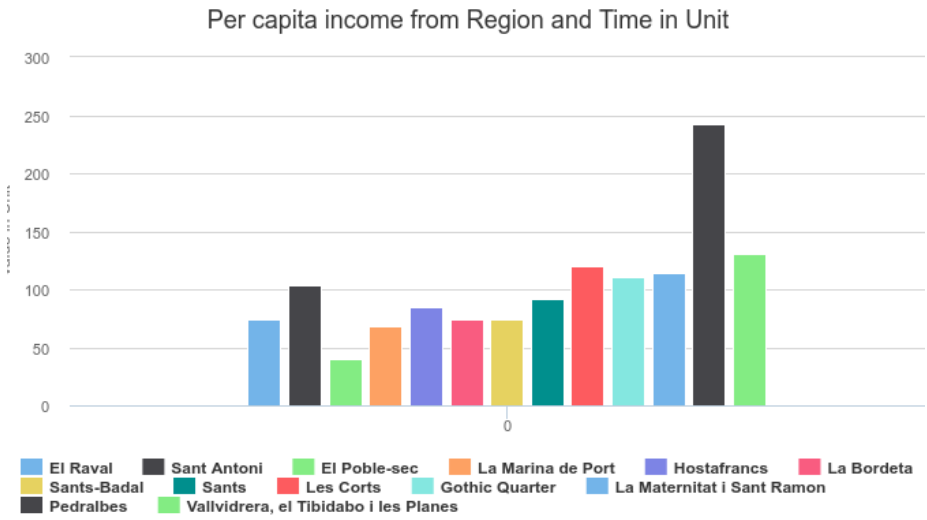

Figure 8: Per capita income in the *Region* (selected zones) and *Time* (2017/February) dimensions.

Linking to other data sources makes it possible to add further information to enhance the final dataset. Wikidata provides a full set of properties that can be exploited, such as administrative subdivisions, dimensions, images and geographic proximity. Listing 5 shows an example of a SPARQL federated query that is employed to execute queries distributed over different SPARQL endpoints by means of the `SERVICE` keyword. This query was executed from our SPARQL endpoint and merge data from Wikidata (such as geographic coordinates, area occupied by a region and additional external identifiers) as an example of how expert users are allowed to exploit our dataset.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX ex: <http://example.com/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

SELECT DISTINCT ?s ?area ?coordinates ?osmid
WHERE {
  ?s a ex:Region .
```

```
  ?s owl:sameAs ?wd .
  SERVICE <https://query.wikidata.org/sparql>
  {
    ?wd wdt:P625 ?coordinates .
    OPTIONAL {?wd wdt:P2046 ?area . }
    OPTIONAL {?wd wdt:P402 ?osmid . }
  }
}
```

Listing 5: SPARQL federated query retrieving additional information such as the area of the location, the OSM identifier and the geographic coordinates at Wikidata.

5. Discussion and conclusions

The publication of Open Data has attracted the interest of the research community to a great extent. Open data may be visually available online, but are generally of poor quality, thus discouraging others from contributing to and reusing that data. In this paper we have defined a framework suitable for publishing and exploiting linked data including a new step of LOD assessment. The framework uses Semantic Web standards published by the W3C, such as RDF and SPARQL, and focuses mainly on providing and facilitating the analysis of multidimensional models. The main motivation for our research is based on how to increase the value of Open Data and make it useful enriching and assessing the quality of the data before exploitation.

The proposed framework consists of 6 steps: (1) *data source specification* (integration of different data through the ETL process in order to normalise the data obtained from heterogeneous data sources); (2) *RDF data modelling* (transformation of the original sources into the form of a multidimensional data model based on RDF Data Cube vocabulary); (3) *data generation* (the data is then used to generate the RDF data by means of OpenRefine); (4) *data publishing* (the RDF data is enriched and stored in our repository); (5) *data assessment* (the methodology to evaluate the final dataset is provided); (6)

data exploitation (dashboards and a public SPARQL endpoint are provided to interact with the dataset).

The main contributions of our framework are the following:

- The enrichment of the original dataset by using links to external repositories. In our experimentation, we established links to Wikidata and GeoNames. In addition, data enrichment allows the inclusion of new indicators by means of federated SPARQL queries.

- A methodology to evaluate a dataset based on the RDF Data Cube vocabulary. To solve the problem of the evaluation of the data quality we decided to adapt the measures that are commonly used in the knowledge graph domain to the specificities of data cube repositories. The original methodology includes some criteria which may not be applicable to evaluate the LOD created from Open Data platforms such as the criterion consistency of schema restrictions during the insertion of new statements.

- The dataset exploitation using: (1) dashboards that allow non-expert users to interact with data generated according to the RDF Data Cube vocabulary; (2) a public SPARQL endpoint for expert users.

The approach has been evaluated by using the data from the Barcelona Open Data platform, and specifically, the state of critical cleaning spots in the city of Barcelona.

We foresee several opportunities to improve our work, such as including the data from other cities and adding multiple Linked Data repositories. We also plan to improve our dataset by using more vocabularies in addition to evaluating new methods for the visualisation and exploitation of Government Linked Data. Finally, the results of the data quality process will be taken into account in order to identify features to improve the publication of the dataset such as the addition of provenance and ranking information.

Acknowledgements

This work was supported in part by the Spanish Ministry of Science, Innovation and Universities through the Project ECLIPSE-UA under Grant RTI2018-094283-B-C32.

References

[1] Miniwatts Marketing Group, World Internet Users and 2018 Population Stats, https://www.internetworldstats.com/stats.htm, [Online; accessed 4-April-2018] (2018).

[2] C. K. Emani, N. Cullot, C. Nicolle, Understandable big data: A survey, Computer Science Review 17 (2015) 70–81. doi:10.1016/j.cosrev.2015.05.002.
URL https://doi.org/10.1016/j.cosrev.2015.05.002

[3] J. Marden, C. Li-Madeo, N. Whysel, J. Edelstein, Linked open data for cultural heritage: evolution of an information technology, in: M. J. Albers, K. Gossett (Eds.), Proceedings of the 31st ACM international conference on Design of communication, Greenville, NC, USA, September 30 - October 1, 2013, ACM, 2013, pp. 107–112. doi:10.1145/2507065.2507103.
URL https://doi.org/10.1145/2507065.2507103

[4] Tim Berners-Lee, Linked Data, https://www.w3.org/DesignIssues/LinkedData.html, [Online; accessed 5-April-2018] (2006).

[5] RDF Working Group, Resource Description Framework (RDF), http://www.w3.org/RDF, [Online; accessed 15-November-2018] (2014).

[6] S. Scheider, A. Degbelo, R. Lemmens, C. van Elzakker, P. Zimmerhof, N. Kostic, J. Jones, G. Banhatti, Exploratory querying of SPARQL endpoints in space and time, Semantic Web 8 (1) (2017) 65–86. doi:10.3233/SW-150211.
URL https://doi.org/10.3233/SW-150211

[7] Madrid City Council, Portal de datos abiertos del Ayuntamiento de Madrid, https://datos.madrid.es/, [Online; accessed 5-April-2018] (2018).

[8] London City Council, London Datastore, https://data.london.gov.uk/, [Online; accessed 5-April-2018] (2018).

[9] Barcelone City Council, Open Data BCN, http://opendata-ajuntament.barcelona.cat/en/, [Online; accessed 5-April-2018] (2018).

[10] Paris City Council, Open Data Paris, https://opendata.paris.fr, [Online; accessed 5-April-2018] (2018).

[11] New York City Council, NYC Open Data, https://opendata.cityofnewyork.us/, [Online; accessed 5-April-2018] (2018).

[12] G. J. P. Link, K. Lumbard, K. Conboy, M. Feldman, J. Feller, J. George, M. Germonprez, S. P. Goggins, D. Jeske, G. Kiely, K. Schuster, M. Willis, Contemporary issues of open data in information systems research: Considerations and recommendations, CAIS 41 (2017) 25.
URL http://aisel.aisnet.org/cais/vol41/iss1/25

[13] E. Ruijer, S. Grimmelikhuijsen, M. J. Hogan, S. Enzerink, A. Ojo, A. Meijer, Connecting societal issues, users and data. scenario-based design of open data platforms, Government Information Quarterly 34 (3) (2017) 470–480. doi:10.1016/j.giq.2017.06.003.
URL https://doi.org/10.1016/j.giq.2017.06.003

[14] F. Benitez-Paez, A. Degbelo, S. Trilles, J. Huerta, Roadblocks hindering the reuse of open geodata in colombia and spain: A data user's perspective, ISPRS Int. J. Geo-Information 7 (1) (2018) 6. doi:10.3390/ijgi7010006.
URL https://doi.org/10.3390/ijgi7010006

[15] S. Bogdanović-Dinić, N. Veljković, L. Stoimenov, How Open Are Public Government Data? An Assessment of Seven Open Data Portals, Springer New York, New York, NY, 2014, pp. 25–44. doi:10.1007/

978-1-4614-9982-4_3.
URL https://doi.org/10.1007/978-1-4614-9982-4_3

[16] C. Muller, L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, R. Leigh, Crowdsourcing for climate and atmospheric sciences: current status and future potential, International Journal of Climatology 35 (11) (2015) 3185–3203. arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4210, doi:10.1002/joc.4210.
URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4210

[17] Open Data Barometer, Global report, https://opendatabarometer.org/4thedition/report/, [Online; accessed 5-April-2018] (2018).

[18] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, Dbpedia: A nucleus for a web of open data, in: K. Aberer, K. Choi, N. F. Noy, D. Allemang, K. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007., Vol. 4825 of Lecture Notes in Computer Science, Springer, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.
URL https://doi.org/10.1007/978-3-540-76298-0_52

[19] T. P. Tanon, D. Vrandecic, S. Schaffert, T. Steiner, L. Pintscher, From freebase to wikidata: The great migration, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 1419–1428. doi:10.1145/2872427.2874809.
URL https://doi.org/10.1145/2872427.2874809

[20] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, G. Weikum,

YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames, in: P. T. Groth, E. Simperl, A. J. G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck, Y. Gil (Eds.), The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II, Vol. 9982 of Lecture Notes in Computer Science, 2016, pp. 177–185. doi:10.1007/978-3-319-46547-0_19.
URL https://doi.org/10.1007/978-3-319-46547-0_19

[21] The World Bank, Starting an Open Data Initiative, http://opendatatoolkit.worldbank.org/en/starting.html, [Online; accessed 5-April-2018] (2013).

[22] European Data Portal, Open Data in a nutshell, https://www.europeandataportal.eu/en/providing-data/goldbook/open-data-nutshell, [Online; accessed 5-April-2018] (2015).

[23] A. K. Ojo, E. Curry, F. A. Zeleti, A tale of open data innovations in five smart cities, in: T. X. Bui, R. H. S. Jr. (Eds.), 48th Hawaii International Conference on System Sciences, HICSS 2015, Kauai, Hawaii, USA, January 5-8, 2015, IEEE Computer Society, 2015, pp. 2326–2335. doi:10.1109/HICSS.2015.280.
URL https://doi.org/10.1109/HICSS.2015.280

[24] H. Dong, G. Singh, A. Attri, A. El-Saddik, Open data-set of seven canadian cities, IEEE Access 5 (2017) 529–543. doi:10.1109/ACCESS.2016.2645658.
URL https://doi.org/10.1109/ACCESS.2016.2645658

[25] European Data Portal, Some insights in the most popular open data categories, https://www.europeandataportal.eu/en/highlights/some-insights-most-popular-open-data-categories, [Online; accessed 2-November-2018] (2017).

[26] Open Knowledge International, Open Data Handbook: File formats, http:

//opendatahandbook.org/guide/en/appendices/file-formats/, [Online; accessed 7-April-2018] (2012).

[27] A. Sourouni, G. Kourlimpinis, S. Mouzakitis, D. Askounis, Towards the government transformation: An ontology-based government knowledge repository, Computer Standards & Interfaces 32 (1-2) (2010) 44–53. doi:10.1016/j.csi.2009.06.002.
URL https://doi.org/10.1016/j.csi.2009.06.002

[28] A. Nogales, M. Sicilia, S. S. Alonso, E. G. Barriocanal, Linking from schema.org microdata to the web of linked data: An empirical assessment, Computer Standards & Interfaces 45 (2016) 90–99. doi:10.1016/j.csi.2015.12.003.
URL https://doi.org/10.1016/j.csi.2015.12.003

[29] D. Bianchini, V. D. Antonellis, M. Garda, M. Melchiori, Exploiting smart city ontology and citizens' profiles for urban data exploration, in: H. Panetto, C. Debruyne, H. A. Proper, C. A. Ardagna, D. Roman, R. Meersman (Eds.), On the Move to Meaningful Internet Systems. OTM 2018 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I, Vol. 11229 of Lecture Notes in Computer Science, Springer, 2018, pp. 372–389. doi:10.1007/978-3-030-02610-3_21.
URL https://doi.org/10.1007/978-3-030-02610-3_21

[30] M. Rani, S. Alekh, A. Bhardwaj, A. Gupta, O. P. Vyas, Ontology-based classification and analysis of non- emergency smart-city events, CoRR abs/1708.00856. arXiv:1708.00856.
URL http://arxiv.org/abs/1708.00856

[31] World Wide Web Consortium (W3C), Data on the Web Best Practices, https://www.w3.org/TR/dwbp/, [Online; accessed 19-June-2018] (2017).

[32] E. D. Portal, Re-using Open Data, https://www.europeandataportal.

eu/sites/default/files/re-using_open_data.pdf, [Online; accessed 9-April-2018] (2017).

[33] A. Piscopo, Wikidata:Requests for comment/Data quality framework for Wikidata, https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata, [Online; accessed 11-February-2018] (2016).

[34] M. Färber, F. Bartscherer, C. Menne, A. Rettinger, Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO, Semantic Web 9 (1) (2018) 77–129. doi:10.3233/SW-170275.
URL https://doi.org/10.3233/SW-170275

[35] Y. Lee, A life-cycle workflow architecture for linked data, in: Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, ICMLSC 2017, Ho Chi Minh City, Vietnam, January 13-16, 2017, 2017, pp. 117–121. doi:10.1145/3036290.3036302.
URL https://doi.org/10.1145/3036290.3036302

[36] M. Lnenicka, J. Komarkova, Developing a government enterprise architecture framework to support the requirements of big and open linked data with the use of cloud computing, Int J. Information Management 46 (2019) 124–141. doi:10.1016/j.ijinfomgt.2018.12.003.
URL https://doi.org/10.1016/j.ijinfomgt.2018.12.003

[37] Working Group Note, Best Practices for Publishing Linked Data, https://www.w3.org/TR/ld-bp/, [Online; accessed 20-May-2018] (2014).

[38] D. W. Bernadette Hyland, The joy of data - cookbook for publishing linked government data on the web., http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook (2011).

[39] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, A. Gómez-Pérez, Methodological Guidelines for Publishing Government Linked Data, Springer New York, New York, NY, 2011, Ch. –, pp. 27–49. doi:

10.1007/978-1-4614-1767-5_2.
URL https://doi.org/10.1007/978-1-4614-1767-5_2

[40] S. Hira, P. Deshpande, Data analysis using multidimensional modeling, statistical analysis and data mining on agriculture parameters, Procedia Computer Science 54 (2015) 431 – 439, eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India. doi:https://doi.org/10.1016/j.procs.2015.06.050.
URL http://www.sciencedirect.com/science/article/pii/S1877050915013745

[41] M. H. Carrasco, S. Luján-Mora, A. Maté, Evaluating open access journals using semantic web technologies and scorecards, J. Information Science 43 (1) (2017) 3–16. doi:10.1177/0165551515624353.
URL https://doi.org/10.1177/0165551515624353

[42] R. Cyganiak, D. Reynolds, The RDF Data Cube Vocabulary, https://www.w3.org/TR/vocab-data-cube/, [Online; accessed 8-April-2018] (2014).

[43] E. Kalampokis, E. Tambouris, K. A. Tarabanis, Linked open government data analytics, in: M. Wimmer, M. Janssen, H. J. Scholl (Eds.), Electronic Government - 12th IFIP WG 8.5 International Conference, EGOV 2013, Koblenz, Germany, September 16-19, 2013. Proceedings, Vol. 8074 of Lecture Notes in Computer Science, Springer, 2013, pp. 99–110. doi:10.1007/978-3-642-40358-3_9.
URL https://doi.org/10.1007/978-3-642-40358-3_9

[44] L. Etcheverry, A. A. Vaisman, QB4OLAP: A vocabulary for OLAP cubes on the semantic web, in: J. F. Sequeda, A. Harth, O. Hartig (Eds.), Pro-

ceedings of the Third International Workshop on Consuming Linked Data, COLD 2012, Boston, MA, USA, November 12, 2012, Vol. 905 of CEUR Workshop Proceedings, CEUR-WS.org, 2012, pp. –.
URL http://ceur-ws.org/Vol-905/EtcheverryAndVaisman_COLD2012.pdf

[45] L. Etcheverry, S. A. Gómez, A. A. Vaisman, Modeling and querying data cubes on the semantic web, CoRR abs/1512.06080. arXiv:1512.06080.
URL http://arxiv.org/abs/1512.06080

[46] J. Matsuda, A. Mizutani, Y. Asano, D. Yamamoto, H. Takeda, I. Ohmukai, F. Kato, S. Koide, H. Harada, S. Nishimura, Publication of statistical linked open data in japan, in: Semantic Technology - 8th Joint International Conference, JIST 2018, Awaji, Japan, November 26-28, 2018, Proceedings, 2018, pp. 307–319. doi:10.1007/978-3-030-04284-4_21.
URL https://doi.org/10.1007/978-3-030-04284-4_21

[47] J. Klímek, J. Kucera, M. Necaský, D. Chlapek, Publication and usage of official czech pension statistics linked open data, J. Web Semant. 48 (2018) 1–21. doi:10.1016/j.websem.2017.09.002.
URL https://doi.org/10.1016/j.websem.2017.09.002

[48] L. Galárraga, K. A. M. Mathiassen, K. Hose, Qboairbase: The european air quality database as an RDF cube, in: Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017., 2017.
URL http://ceur-ws.org/Vol-1963/paper507.pdf

[49] K. Abicht, G. Alkhouri, N. Arndt, R. Meissner, M. Martin, Cubeviz.js: A lightweight framework for discovering and visualizing RDF data cubes, in: M. Eibl, M. Gaedke (Eds.), 47. Jahrestagung der Gesellschaft für Informatik, Informatik 2017, Chemnitz, Germany, September 25-29, 2017, Vol.

P-275 of LNI, GI, 2017, pp. 1915–1921. doi:10.18420/in2017_191.
URL https://doi.org/10.18420/in2017_191

[50] E. Folmer, W. Beek, L. Rietveld, S. Ronzhin, R. Geerling, D. den Haan, Enhancing the usefulness of open governmental data with linked data viewing techniques, in: 52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019, 2019, pp. 1–10.
URL http://hdl.handle.net/10125/59728

[51] E. Kalampokis, D. Zeginis, K. A. Tarabanis, On modeling linked open statistical data, J. Web Semant. 55 (2019) 56–68. doi:10.1016/j.websem.2018.11.002.
URL https://doi.org/10.1016/j.websem.2018.11.002

[52] G. Candela, P. Escobar, R. C. Carrasco, M. Marco-Such, A linked open data framework to enhance the discoverability and impact of culture heritage, Journal of Information Science 0 (0) (0) 0165551518812658. arXiv:https://doi.org/10.1177/0165551518812658, doi:10.1177/0165551518812658.
URL https://doi.org/10.1177/0165551518812658

[53] S. K. Bansal, Towards a semantic extract-transform-load (ETL) framework for big data integration, in: 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014, IEEE Computer Society, 2014, pp. 522–529. doi:10.1109/BigData.Congress.2014.82.
URL https://doi.org/10.1109/BigData.Congress.2014.82

[54] R. Kimball, The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, John Wiley, 1996.

[55] E. Acheson, S. D. Sabbata, R. S. Purves, A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, Computers, Environment and Urban Systems 64 (2017) 309–320. doi:10.1016/j.

compenvurbsys.2017.03.007.
URL https://doi.org/10.1016/j.compenvurbsys.2017.03.007

[56] Z. Pan, T. Zhu, H. Liu, H. Ning, A survey of RDF management technologies and benchmark datasets, J. Ambient Intelligence and Humanized Computing 9 (5) (2018) 1693–1704. doi:10.1007/s12652-018-0876-2.
URL https://doi.org/10.1007/s12652-018-0876-2

[57] D. Faye, O. Curé, G. Blin, A survey of rdf storage approaches, ARIMA Journal 15 (2012) 11–35.

[58] World Wide Web Consortium (W3C), Comparison of rdfjs libraries, https://www.w3.org/community/rdfjs/wiki/Comparison_of_RDFJS_libraries, [Online; accessed 29-November-2018] (2018).

[59] World Wide Web Consortium (W3C), Describing linked datasets with the void vocabulary, https://www.w3.org/TR/void/, [Online; accessed 19-June-2018] (2011).

[60] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, J. of Management Information Systems 12 (4) (1996) 5–33.
URL http://www.jmis-web.org/articles/1002

[61] W3C, Notation3 (n3): A readable rdf syntax, "https://www.w3.org/TeamSubmission/n3/", [Online; accessed 13-November-2018] (2011).