Are the Methodologies for Producing Linked Open Data Feasible for Public Administrations?

Roberto Boselli, Mirko Cesarini, Fabio Mercorio and Mario Mezzanzanica
Department of Statistics and Quantitative Method, CRISP Research Centre, University of Milan Bicocca, Milan, Italy

Keywords: Linked Open Data, Ontologies, Open Data, Public Administration Information.

Abstract: Linked Open Data (LOD) enable the semantic interoperability of Public Administration (PA) information. Moreover, they allow citizens to reuse public information for creating new services and applications. Although there are many methodologies and guidelines to produce and publish LOD, the PAs still hardly understand and exploit LOD to improve their activities. In this paper we show the use of a set of best practices to support an Italian PA in producing LOD. We show the case of LOD production from existing open datasets related to public services. Together with the production of LOD we present the definition of a reference ontology, the Public Service Ontology, integrated with the datasets. During the application, we highlight and discuss some critical points we found in methodologies and technologies described in the literature, and we identify some potential improvements.

1 INTRODUCTION

Public Administrations (PAs) own an enormous wealth of data and information that, once shared, may be used to produce innovative applications and services able to increase the effectiveness and efficiency of PA activities. According to both the Open Government movement (see, e.g., (Trivellato, 2014)) and the directives of Digital Agenda for Europe[1], PAs are encouraged to publish their data on the Web in an open format. To this end, the Digital Agenda also enacted guidelines to achieve semantic interoperability through Linked Open Data (LOD). If properly implemented, these actions can effectively contribute in exploiting public information and enabling citizens to collaborate with both policy-makers and service providers to improve governance, public life, and public services.

However, it is still not easy for PAs to produce and publish LOD, and this makes hard for non-expert users the reuse of public information to create new services and applications.

In this paper we discuss the application of a set of best practices to support an Italian PA in producing LOD, and we highlight some critical points that can be found in methodologies and technologies described in the literature. Moreover, our aim is to identify some potential improvements in defining a replicable methodology, suitable for non-expert users, and supporting activity automation.

The paper is organized as follows. Section 2 provides an overview of Linked Open Data context, focusing on the motivations and requirements for PAs to use them. Section 3 presents the methodological aspects and related work from the literature discussed to define the best practices in producing LOD. Section 4 shows the application domain and discusses some points addressed during the implementation and use of LOD technologies and principles. Section 5 concludes the paper with an outline of our on-going research directions.

[1] http://ec.europa.eu/digital-agenda/

2 LINKED OPEN DATA

Linked Data refer to a set of best practices based on the following four principles (Berners-Lee, 2006):

1. use URIs as names for things;
2. use HTTP URIs so those names can be looked up (dereferencing);
3. provide useful information upon lookup of those URIs (using the standards RDF (Manola, 2004) and SPARQL (Harris, 2010));

Boselli R., Cesarini M., Mercorio F. and Mezzanzanica M..
Are the Methodologies for Producing Linked Open Data Feasible for Public Administrations?.
DOI: 10.5220/0005143303990407
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (KomIS-2014), pages 399-407
ISBN: 978-989-758-035-2

4. include links to other URIs to discover new knowledge.

Linked Data can be seen as a bottom-up approach to Semantic Web adoption (Bizer, 2009): the aforementioned four principles represent guidelines for collaboratively publishing and interlinking structured data over the Web by exploiting Semantic Web standards.

Moreover, the W3C consortium published further technical guidelines (W3C, 2009) to promote the adoption of Linked Data approach by PAs and to encourage them to publish public data in the LOD format. These guidelines are summarized in the five-stars Linked Data scheme for Web publishing (Berners-Lee, 2006):

- 1 star: Publish under an open licence;
- 2 stars: Publish machine-readable structured data;
- 3 stars: Use non-proprietary formats;
- 4 stars: Use URIs and open standards (RDF and SPARQL) to identify things;
- 5 stars: Link your data to other data.

In addition to the schema, Berners-Lee explains why the Linked Data are the best way to meet the three main requirements for which government data should be made available on Web (Berners-Lee, 2009):

1. to increase citizen awareness of government functions to enable greater accountability;
2. to contribute valuable information about the world;
3. to enable the government, the country, and the world to function more efficiently.

This approach produced the LOD Cloud (Bizer, 2007a), a graph representing the amount of interconnected data published by public and private bodies. In September 2011 LOD came to more than 31 billion RDF triples and more than 500 million RDF links (Jentzsch, 2011). The LOD Cloud provides an ideal environment facilitating the interoperability between datasets. The possibilities are endless when one considers the large amount of LOD already available, for example DBpedia[2], Geonames[3], WordNet[4], the DBLP[5] bibliography etc. One of the best well-known initiative exploiting the LOD Cloud links is the BBC Music site[6]. It gathers music data from several LOD datasets, e.g. MusicBrainz, LastFM, DiscoGS, and mash up them with the BBC heritage to provide an hypertext navigation of music information.

Focusing on the domains covered by triples in the LOD Cloud, the most of the triples are relative to PAs (Government) data, followed by geographical and cross-domain data.

In the next section we will clarify why a PA should publish datasets in the LOD format.

2.1 Motivation

The Linked Data community vision is very simple: to transform the Web into an open and interoperable environment where data are not sealed in independent silos, but interrelated among them.

According to this philosophy, government data have to become of public interest, so that people and applications can access and interpret data using common Web technologies. The Linked Data philosophy does not describe new further technologies and languages (e.g., w.r.t. those proposed by the Semantic Web), but it proposes the rules to follow to make available and accessible information on the Web to both humans and software applications. The aim is to define information assets managed by PAs in a shared and semantically meaningful way.

According to the LOD approach, PAs can publish datasets and emphasize links with other public datasets in the LOD Cloud. This provides a universal access to such data, and it also enables LOD to become the basis of a new paradigm of applications and services design and development.

Once published, public data significantly increase their cognitive value as different datasets, produced and published independently by various parties, can be freely crossed by third parties. To achieve this goal it is necessary that an active collaboration arises between different PAs, organizations and citizens. It is fundamental that to develop new applications based on different LOD datasets, these share a common language and semantics fostering unique interpretations. This is achieved by using the Semantic Web languages, tools and standards, in particular the use of ontologies and shared vocabularies.

Nevertheless, when humans or applications try to combine data from heterogeneous sources and domains, a conceptual description is required to guarantee semantics to the data. Structured data in the LOD Cloud can be managed from two points of view, namely *instance level* and *schema (ontological) level* (Jain, 2010b). The LOD Cloud datasets are mainly inter-linked at the instance level,

[2] http://dbpedia.org
[3] http://www.geonames.org/
[4] http://wordnet.princeton.edu/
[5] http://www.informatik.uni-trier.de/~ley/db/
[6] http://www.bbc.co.uk/music

but they lack schema level mappings, because do not express relationships between concepts of different datasets at the schema level. An improvement of the schema level may be acquired by fostering conceptual descriptions through ontologies.

Therefore, the two main components of this scenario, data and vocabularies (ontologies), should be jointly managed to provide a complete and consistent knowledge base to application developers.

To this aim, an effective methodology should address ontology creation, conversion of structured data from csv (3 stars) into RDF (4 stars) and LOD (5 stars), and perform additional tasks of data quality analysis.

3 METHODOLOGIES TO PRODUCE AND PUBLISH LOD

Our goal is to support PAs in producing and publishing LOD, therefore we discuss and show the application of some best practices identified in the literature. Above all, our goal is to identify some points for potential improvement in defining a methodology that we would like to make as easy and automatized as possible, replicable, and usable by non-expert users, also with a high rate of cost savings.

We use open source tools supported by active and collaborative communities to meet the PA' economization needs. Some tools chosen are de-facto standards into the Linked Data community, such as the ontology editor Protégé[7] or the server Open Link Virtuoso[8] to access data via SPARQL endpoints.

A standard methodology does not exist in the literature as different steps are required depending on the format and quality of the data to publish. Thus, we studied some literature works where the main technical steps are recommended (Bizer, 2007b; Heath, 2011; Lebo, 2011).

An important work for this paper is (Hogan, 2012) as it gives a rich list of guidelines to publish LOD and also a quantitative analysis of the state of the LOD Cloud. Moreover, it provides a literature overview on the main issues affecting the Web of LOD, some of these we discuss in what follows.

If the data come from relational databases, data manipulation techniques are needed for converting the database content into its RDF representation (i.e. the RDF schema). The conversion requires a mapping process between the database and the RDF schema. Since each conversion provides a new interpretation of the original data, ontologies or vocabularies are used to define the meaning of the entities involved in the mapping, namely: *classes*, *properties* and *instances*.

Unfortunately, ontologies are not available for all LOD domains, and this often requires the definition of a new ontology. Then, the new ontology has to be integrated with both the produced RDF schemas and the other datasets already available in the LOD Cloud.

The integration task is a critical aspect of all the methodologies, because a human supervision is required to assess the ontology mappings. However, tools and methods are currently available to identify the best method to align ontologies and to link the datasets (OAEI; Euzenat 2013).

The works of (Jain, 2010a and 2010b; Parundekar, 2010) on ontology alignment in LOD have been an important reference for us. They explain that links between LOD datasets are almost exclusively on the level of instances and not on schema level. During the experimentation, we were able to confirm the lack of ontological level in the LOD Cloud and the need to provide a reference ontology to the published datasets.

One of the methods most commonly used to integrate different LOD datasets is to annotate data using the *owl:sameAs* property. This method provides declarative semantics for aggregating distributed data, i.e., machines can merge resource descriptions if the resources described are linked with *owl:sameAs*.

However, more researchers and developers agree that the use of *owl:sameAs* does not always integrate "equivalent" resources. It links mainly instances, and often this equivalence is context-dependent (Halpin, 2010; Ding, 2010). To give a few examples, let us consider the population of Milan, one of the most important city in the north of Italian. Focusing on the resident population value, two different results can be obtained from DBpedia (1.350.267 inhabitans) and Geonames (1.236.837 inhabitans). Each value could be true in a certain context, but in answering a simple query a person should expect just one value rather than a set of alternatives. In such a scenario, this clearly represents an issue to be addressed and properly handled. To this end, we believe the use of probabilistic, fuzzy, or statistical approaches to derive *owl:sameAs* links between datasets is a promising approach, as it might provide

[7] http://protege.stanford.edu/

[8] http://virtuoso.openlinksw.com/

users a ranking of results to choose the best fitting one.

The semantic integration among LOD datasets and ontologies based on automated reasoning tasks is discussed in (Hitzler, 2010; Polleres, 2010). These works discuss the benefits and the challenging issues derived from shared inferences. In particular, we perform inferences to enrich ontologies with new entities derived from the datasets. Nevertheless, inconsistencies can occur due to undefined classes and properties. To deal with this issue we started working with data cleaning techniques and it represents an issue to in depth investigate in the future.

Another critical point emerging from the literature that we address in this paper is the difficulty, especially for non-expert users, to analyse and navigate data in the LOD Cloud using SPARQL queries. SPARQL is the de-facto standard query and protocol language to access LOD, but the SPARQL syntax requires users to specify the precise details of the structure of the graph being queried in the triple pattern. Thus, the user has to be familiar with the SPARQL syntax, but in the context of PAs it is not easy to find people with this competence.

At a glance, we define a set of best practices identified from the literature methodologies to publish LOD:

1. building of a domain ontology,
2. conversion of structured data into RDF,
3. generation of RDF schemas of the datasets,
4. integration of ontology and RDF schemas,
5. integration of datasets with the LOD Cloud,
6. implementation of SPARQL endpoints,
7. publication of datasets on the SPARQL endpoint.

The methodologies usually described in the literature drive the creation of a single LOD, but are less effective in integrating different datasets from different domains. For this purpose, our work is based on the definition of a reference ontology for a specific domain (the Public Service Ontology) that is increasingly enriched by the integration of new datasets.

Furthermore, these methodologies are tailored for expert users and most of the works in literature do not provide effective alternatives for people having low technical capabilities. Therefore, we plan to design interfaces and tools facilitating the LOD exploitation and use by PAs.

In the next sections we discuss some potential improvements of the methodologies on the basis of the best practices addressed during the application (discussed below).

4 APPLICATION DOMAIN

Italian PAs are encouraged by a national law to publish their data on the Web in an open format, and several PAs developed Web portals where they publish open data following the W3C guidelines.

The Italian PA which publishes the largest number of datasets is the Lombardy Region through its open data portal[9], with more than 700 datasets published until April 2014. Nevertheless, only the 5% of the Italian open data achieves the 5 stars of Berners-Lee schema[10] at the national level. Among the PAs there is no region that publishes LOD at the current time.

Therefore, the Lombardy administration wish to be the first Italian region in publishing LOD. It is not the only reason that drove the region in this task, another reason is the alignment with the European Interoperability Framework (EIF) to support the European delivery of e-government services (EIF, 2004).

The Lombardy Region and CRISP[11] (Interuniversity Research Centre on Public Services) launched the ITLab project with the scope to develop 5 stars LOD from open datasets. The CRISP research centre develops models, methodologies and tools for collecting, analyzing, and supporting data useful to define and improve services and policies for the public sector.

Since the beginning of the project it was decided to focus on the public services domain. This choice guided the selection of the datasets to use in the application, extracted from Lombardy open data portal. Five datasets related to different public services have been used: schools[12], residential child care institutions[13], nursing houses[14], social cooperatives[15] and hospitals[16].

Public services are delivered to citizens from specialized providers (both public and private) to build complex and comprehensive services (Boselli, 2011). The provision of services is carried out by

[9] https://www.dati.lombardia.it
[10] see statistics available on the National Portal of Italian Open Data, www.dati.gov.it, updated on April 2014.
[11] http://www.crisp-org.it
[12] https://www.dati.lombardia.it/Istruzione/Anagrafe-Scuole/fm99-kxtn
[13] https://www.dati.lombardia.it/Famiglia/Elenco-Comunit-Per-Minori/hs2e-549s
[14] https://www.dati.lombardia.it/Sanit-/Elenco-RSA-Accreditate/vef4-8fnp
[15] https://www.dati.lombardia.it/Solidariet-/Albo-Cooperative-Sociali/tuar-wxya
[16] https://www.dati.lombardia.it/Sanit-/Strutture-di-ricovero-e-cura/teny-wyv8

several actors (belonging to several organizations and having also no direct or hierarchical relationships with public authorities). Some examples are given:

- services for elderly people combining healthcare services (e.g. day-hospital treatment) with public transport (e.g. shuttle busses and steward crew provided by non-profit associations);
- unemployment subsidy (provided by state agencies) granted to unemployed people under the condition that they attend requalification courses (provided by vocational training schools, universities, or professional trainers);
- registry certificates requested by citizens through web forms and delivered by means of courier services.

The selected datasets relate some of the main public service categories: health, family, education and subsidiarity. They are quite different from each other in terms of number of values and attributes: they range from the biggest (schools) with +5000 rows and 20 columns to the smallest (hospitals) with 215 rows. One challenging issue here is the presence of multi-valued cells, e.g. in the social cooperative' dataset the attribute *Target users* which has cells up to 5 different values. The complete set of the datasets is composed by +8600 instances of public service providers geographically located in the Lombardy region.

In order to describe the domain and to provide semantics to datasets a Public Service Ontology has been created, as we describe next, and this task deals with the first point of our best practices.

4.1 Public Service Ontology

In order to make interoperable data published on the Web, it is necessary to refer to shared vocabularies whose semantics is well defined. These shared vocabularies are the ontologies, formal and explicit specifications of a domain conceptualization (Gruber, 1995). Ontologies play a key role in the Semantic Web to facilitate the understanding of the meaning of data by software agents, and this is also true in the Web of Linked Data. Therefore, it is necessary reuse existing ontologies (if any) or to create them from scratch to provide explicit semantics to LOD.

Various ontologies and vocabularies are used to provide valuable knowledge to different domains in the Web of LOD, e.g. Dublic Core, FOAF (Brickley 2010), SIOC (Bojars, 2007), Geonames, DBpedia, UMBEL (Bergman, 2008), and YAGO (Suchanek, 2007). The latter three facilitate data integration across a wide range of interlinked sources, in particular UMBEL and YAGO being upper-level ontologies. Most of these reusable ontologies are written in RDF and OWL, the Semantic Web languages, and can be easily imported during the building of a new one (Boselli, 2012).

Nevertheless, the organizations that publish LOD integrated with ontologies are a little portion of the LOD Cloud publishers. According to some empirical observations and the literature articles (Jain, 2010a and b; Polleres, 2010; Hitzler, 2010) most of data in the LOD Cloud refers to a few ontologies, or parts of ontologies, and so there are few institutions that also develop contextual ontologies in addition to publishing data. In this project we give a contribution in this direction, with the publication of both data and a reference integrated ontology.

This is the Public Service Ontology, created in OWL2 (Hitzler, 2009) by using the ontology editor Protégé. Some vocabularies and conceptual models are used as theoretical reference to describe the public service domain. The first model taken into consideration is the Core Public Service Vocabulary (CPSV) which defines a method to describe a public service (ISA, 2013). A second model considered is the Local Government Business Model, the conceptual basis for the UK online platform Effective Service Delivery Toolkit (ESD-Toolkit)[17].

The ontology we propose is different from the CPSV by the fact that the latter is designed to describe the process at the core of the service rather than the service itself, focusing on inputs and outputs. While the Local Government Business Model is a rich description of tools and models useful to classify local public services on British territory, therefore is strongly context-dependent.

Our ontology is thought to be the most general possible, although describing a specific domain. In the Figure 1 some of the main entities of the Public Service Ontology are represented. The public service concept is described in our ontology as a set of functions and actions carried out by agents who are geographically located. The agents can be service providers or final users. Some parts of the above models are imported for building the new one, while new concepts and properties are defined from scratch. The percentage of reused entities in our ontology is near the 80%, the new concepts added are mainly those derived from the datasets, as we explain in the section 4.2.

[17] http://www.esd.org.uk/

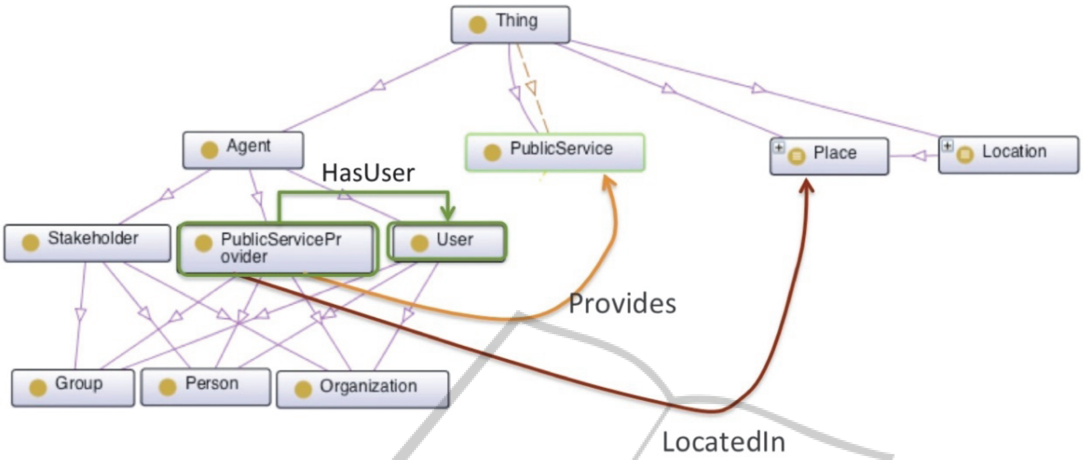

Figure 1: The Public Service Ontology main classes and properties.

Some of the existing concepts imported into the Public Service Ontology are: the CPSV concepts *PublicService*, *Channel*, *Agent*, that overlap with similar concepts of the ESD-Toolkit, while a specific concept of the latter reused in our ontology is *Function*. Furthermore, other ontologies provide us some fundamental concepts, e.g. *Person* concept and *based_near* property from FOAF; *Address* concept from Vcard[18] and *Location* concept from DBpedia.

The ontology development task is carried out in parallel with the conversion of the existing datasets in RDF, and this second task contributes to refine and complete the ontology: the concepts and instances entered into the ontology are identified while the five datasets are converted in RDF, as we describe in the following.

4.2 Conversion of Datasets from csv to LOD

In this section we present a procedure to convert open datasets from 3 stars (csv) to 5 stars (LOD). In the project we used LODRefine, an extension of OpenRefine[19], an open source software tool created with the aim to facilitate the data cleaning procedures. In particular, LODRefine allows two basic operations for manipulating datasets and their transformation into LOD: first, the construction of RDF Schema starting from the structure of the tables, second the semi-automatic reconciliation of values.

[18] http://www.w3.org/2006/vcard/
[19] http://openrefine.org/

The procedure steps that LODRefine supports are the following:

1. Importing the dataset from relational databases,
2. Cleaning of the fields,
3. Reconciliation,
4. RDF schema definition,
5. Exporting RDF schemas.

These steps allow one to accomplish the points 2 and 3 of the best practices discussed in Section 3. We achieved the others (from 4 to 7) by using Protégé and Open Link Virtuoso.

The downloaded datasets from the Lombardy portal are tables with csv data licensed with IODL (Italian Open Data License). LODRefine allows to easily import and visualize the datasets as relational tables. A first processing of the dataset leads to select the columns of interest, the possible creation of new columns that contain information derived from other data, and a general reorganization of the table. Moreover, LODRefine allows one to clean data, e.g. it is possible to apply transformations to the value contained in the cells or the type of data associated with them or split the multi-valued cells. An important function for data quality purposes provided by LODRefine is the cluster analysis, which can be used to unify, standardize and complete the data.

One of the most useful functions of LODRefine that we used is the reconciliation of the URIs, thanks to the presence of the RDF Extension. The basic idea of the URI reconciliation is to recover from an external data source the URIs to be associated with the data in the dataset. This, in effect, allows one to connect the raw data in the table to external data, just by following the philosophy of the LOD. In fact, once the external resource is represented as URI, reconciliation enables to enrich the dataset with URIs that link the dataset to the rest of the LOD Cloud.

In such a case, we choose to use as server for the reconciliation the DBpedia SPARQL endpoint. LODRefine queries the DBpedia endpoint and automatically imports the results in the tables. The main fields on which we apply the reconciliation are those related to the name of the place where the services are located. Each place name in the original dataset is replaced with the corresponding instance of *dbpedia:Place* class. In case of multiple choices, LODRefine requires the user to manually select the right choice. The benefits of this feature are a complete automation of a SPARQL query in a transparent way for the user, and the import of additional information in the original datasets (e.g., by adding directly in the dataset the number of inhabitants or the district of a municipality).

Another important feature made available with the RDF Extension is the design of the RDF schema to associate to each dataset. The purpose of this RDF schema is to define a conceptual model describing the dataset. For this step, the Public Service Ontology provides all the information to represent the dataset with the corresponding RDF schema. By using the ontology as a reference, and with LODRefine, a non-expert user could easily define the RDF schema even without competences in schema modelling.

Each RDF schema can be then exported in Protégé where can be automatically integrated with the ontology, and additional information inferred by the reasoning engine included in Protégé enrich the ontology itself. For example, the integration of the nursing houses' dataset add the *NursingHouse* subclass to the *ServiceProvider* class and also all the instances of that subclass.

At this point, the datasets are in RDF format, linked to external resources (through the reconciliated links to DBpedia resources) and enriched with concepts and properties of both internal and external ontologies. To get the 5 stars, according to the Berners-Lee scale, it is now necessary to publish them on the SPARQL endpoint implemented on Virtuoso.

The final results of these tasks is a set of LOD with +200k RDF triples, 156 classes and 55 properties related to the public service domain available on the Web.

4.3 Discussion on Open Issues

With the procedure described in previous section a PA could produce and publish LOD from its open data. Nevertheless, it is worth discussing some remarks and open issues, both methodological and technological, arisen during the application domain. These open issues affect different aspects of LOD research: usability of LOD technologies, semantic integration of ontologies and datasets, LOD persistence and quality etc. Starting from this, we would open a discussion by suggesting some solutions and actions identified during our work.

Starting from external resources to the dataset it is possible to link to other datasets and retrieve the information they contain. These additional information retrieved enrich the dataset prior to the publication, or retrieve other information by performing federated queries distributed over different SPARQL endpoints. But in some SPARQL endpoints we encountered some bugs related to how strings are managed. We debugged and found, for example, that during the performing of federated queries, if a URI contains accents or tittles (it often happens with Italian names of places) the query does not work or even crash the server.

The best practices described above are applicable to any open dataset of Lombardy portal, be it related to public services or other domain. If the data are not related to public services a different reference ontologies should be used, but the tools that support the logical flow of work don't change. However, the ontology development task remains heavily expert domain-dependent and it is difficult to identify what ontologies and links to import and reuse in datasets. A strong aid would be the identification of reference ontologies or vocabularies to use, in this sense a great support is provided by the good collection of vocabularies available at the Linked Open Vocabularies (LOV) site[20].

By working on the public service datasets we obtain their integration and publication into a single endpoint. But the actual integrated dataset is the result of five examples of public services. In order to arrive to an effective publication and use by end users interested in the Lombardy public services, we should integrate many other datasets of the same domain. It requires an enrichment of the Public Service Ontology, partially performed with the inferences, and the availability of additional data. To do this, the linkage to external resources may be

[20] http://lov.okfn.org/dataset/lov/

improved with the use of some tools helping to discover relationships between different datasets (e.g. Silk[21]) (Volz, 2009), to automatically identify concepts and properties in the LOD Cloud to import in our knowledge base.

The publication of data on an endpoint is the conclusion of the process, but it does not guarantee the use of the data. The publication on the endpoint limits its use to professionals with specific expertise in SPARQL. The use of LOD by an audience of non-experts can only be guaranteed by the introduction of a tool that makes access to data hiding the SPARQL queries. Moreover, user-friendly interfaces to query datasets and to facilitate the visualization of results are needed.

It must be emphasized that an evaluation of the best practices is needed in comparison with others, and to apply it at large scale with larger datasets and ontologies. Moreover, we plan to test and evaluate the benefits of using them in terms of time, number of interactions with the tools for manually matching/cleaning, etc.

5 CONCLUSIONS

The research on Linked Open Data is very rich of contributions and open challenges. In this paper we wanted to give our initial contribution by discussing the methodologies presented in the literature for producing and publishing LOD, according to our experience in this field. Furthermore, the application of one of them to a real domain allowed us to identify some points on which to open a further discussion.

Some steps of the best practices we used are already consolidated and shared in the Linked Data community, while others need a global assessment. Among these, we include the usability of LOD technologies, the semantic integration of ontologies and datasets, the LOD persistence and quality issues. The latter is not discussed in this paper but in our future works we plan to deeply address the quality issues of LOD.

Regarding the development and implementation of PA's LOD, we plan to design and develop a tool with user-friendly interfaces to navigate the SPARQL results, and integrate this tool within an Information System supporting the provision of public sector services.

[21] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

REFERENCES

Bergman, M. K., Giasson, F., 2008. *UMBEL ontology. technical documentation*. Technical report, Structured Dynamics LLC.

Berners-Lee, T., 2006. Linked Data. *W3C Design Issues*.

Berners-Lee, T., 2009. Putting Government Data online. *W3C Design Issues*.

Bizer, C., Heath, T., Ayers, D., Raimond, Y., 2007a. Interlinking Open Data on the Web. In *Demonstrations Track, 4th European Semantic Web Conference*, Innsbruck, Austria.

Bizer, C., Cyganiak, R., Heath, T., 2007b. How to publish linked data on the web. *linkeddata.org Tutorial*.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data – The Story so far. *International Journal on Semantic Web and Information Systems*, 5(3), pp. 1–22.

Bojars, U., Breslin, J. G., Berrueta, D., Brickley, D., Decker, S., Fernández, S., ... Sintek, M., 2007. *SIOC core ontology specification*. WC Member.

Boselli, R., Cesarini, M., Mezzanzanica, M., 2011. Public Service Intelligence: evaluating how the Public Sector can exploit Decision Support Systems. In *Productivity of Services NextGen – Beyond Output / Input. Proceedings of XXI RESER Conference*, Fraunhofer Verlag, Hamburg, Germany.

Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2012. Domain Ontologies Modeling via Semantic Annotations of Unstructured Wiki Knowledge, *Proceedings of the Iadis International Conference (WWW/Internet 2012)*, Madrid, Spain.

Brickley, D., Miller, L. 2010. *FOAF Vocabulary Specification 0.97*.

Ding, L., Shinavier, J., Finin, T., McGuinness, D. L., 2010. owl: sameAs and Linked Data: An empirical study. *WebSci10: Extending the Frontiers of Society On-Line*.

EIF, 2004. *European Interoperability Framework*, White Paper, Brussels.

Euzenat, J., Shvaiko, P., 2013. *Ontology Matching*. Heidelberg: Springer.

Gruber, T. R., 1995. Toward principles for the design of ontologies used for knowledge sharing?. *International journal of human-computer studies*, 43(5), pp. 907-928.

Halpin, H., Hayes, P. J., 2010. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *Proceedings of the WWW2010 workshop on Linked Data on the Web, LDOW2010*.

Harris, S., Seaborne, A., (eds.) 2010. SPARQL Query Language 1.1. *W3C Working Draft*.

Heath, T., Bizer, C., 2011. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, pp.1-136. Morgan & Claypool.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., Rudolph, S. (eds.), 2009. OWL2 Web Ontology Language primer. *W3C Recommendation*.

Hitzler, P., van Harmelen, F. 2010. A Reasonable Semantic Web. *Semantic Web*, 1(1), pp. 39-44.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S., 2012. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14, pp. 14-44.

ISA, Interability solutions for european public administration, 2013. *Core Public Service Vocabulary specification*.

Jain, P., Hitzler, P., Sheth, A. P., Verma, K., Yeh, P. Z. 2010a. Ontology alignment for linked open data. In *The Semantic Web–ISWC 2010,* pp. 402-417. Springer Berlin Heidelberg.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., Sheth, A. P. 2010b. Linked Data Is Merely More Data. In Brickley, D., Chaudhri, V.K., Halpin, H., McGuinness, D., editors, *Linked Data Meets Artificial Intelligence*, pp. 82-86, AAAI Press, Menlo Park, CA.

Jentzsch, A., Cyganiak, R., Bizer, C., 2011. *State of the LOD Cloud*, Version 0.3, http://lod-cloud.net/state/.

Lebo, T., Erickson, J. S., Ding, L., Graves, A., Williams, G. T., DiFranzo, D., ... Hendler, J. 2011. Producing and using linked open government data in the twc logd portal. In *Linking Government Data*, pp. 51-72, Springer New York.

Manola, F., Miller, E., (eds), 2004. *Resource Description Framework (RDF). Primer*. W3C Recommendation.

OAEI the Ontology Alignment Evaluation Initiative: http://oaei.ontologymatching.org/

Parundekar, R., Knoblock, C. A., Ambite, J. L., 2010. Linking and building ontologies of linked data. In *The Semantic Web–ISWC 2010,* Springer Berlin Heidelberg, pp. 598-614.

Polleres, A., Hogan, A., Harth, A., Decker, S. 2010. Can we ever catch up with the Web?. *Semantic Web*, 1(1), pp. 45-52.

Suchanek, F. M., Kasneci, G., & Weikum, G., 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pp. 697-706, ACM.

Trivellato, B., Boselli, R. Cavenago, D., 2014. Design and Implementation of Open Government Initatives at the sub-national Level: Lessons from Italian Cases, in Gascó Hernández, M., (eds.) *Open Government. Opportunities and Challenges for Public Governance*, Springer, pp. 65-84.

Volz, J., Bizer, C., Gaedke, M., Kobilarov, G., 2009. Discovering and maintaining links on theWeb of data. In *ISWC2009*, vol. 5823 of LNCS, p. 650–665, Springer.

W3C, Publishing Open Government data, Working draft 8 september 2009, http://www.w3.org/TR/gov-data/.