From Open Data to Linked Open Data

The GIOCOnDa LOD platform

Lorenzo Sommaruga, Nadia Catenazzi, Davide Bertacco, Riccardo Mazza
Department of Innovative Technologies
University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
CH-6962 Lugano, Switzerland
e-mail: lorenzo.sommaruga@supsi.ch
nadia.catenazzi@supsi.ch
davide.bertacco@supsi.ch
riccardo.mazza@supsi.ch

***Abstract*—In the context of the European project "GIOCOnDa", this paper describes the conversion process from Open Data to Linked Open Data (LOD) and its implementation in the GIOCOnDa LOD platform. The platform contains a number of conversion configurations that allow different data sources from a variety of Open Data domains to be converted into LOD, without the need of software programming. In addition, the platform is configurable and extensible, as it enables to define mapping configurations for new datasets.**

Keywords-Linked Open Data (LOD); GIOCOnDa; OntoPia.

I. INTRODUCTION

This paper describes the methodology and the conversion process from Open Data to Linked Open Data (LOD) implemented in the GIOCOnDa LOD platform in the context of the EU Interreg GIOCOnDa project ("Integrated and holistic management of the open data life cycle" [1]).

This project, funded by the Interreg V-A Italy-Switzerland Programme, aims to create value by developing information products based on the re-use of public Open Data. The project involves the creation of a federated platform for the publication of Linked Open Data by public administrations. In the GIOCOnDa LOD platform, open data, coming from various sources and in different formats, are converted into homogeneous Linked Open Data according to standard ontologies and published together with their metadata. The platform allows conversion of existing 3* Open Data to 5* Open Data, according to the well-known 5-star deployment scheme [2]: data are formalized in RDF, identified by URI and linked to other datasets.

The project focuses on data from the Insubric area, a cross-border territory and community across Italy and Switzerland. According to the project specifications, data useful in the touristic sector are considered. These include data about museums, accommodation facilities and environmental data. The main data sources currently used include: Regione Lombardia open data portal [3] and ARPA (Regional Agency for the Protection of the Environment) [4] for Italian data; Wikidata, Ticino Turismo [5] and OASI [6] for Swiss data.

The GIOCOnDa LOD platform is mainly oriented to domain and ontology experts, who need to authenticate to operate in the platform to create and modify datasets. A public portal is also available, where the datasets produced in the LOD platform are made accessible. Public administrations can submit new datasets for conversion into LOD.

This paper is structured as follows: in Section 2 the methodology adopted to publish LOD data is described; Section 3 is focused on the process of conversion of Open Data to LOD, one of the main steps of this methodology; finally, Section 4 clarifies how this process is implemented in the GIOCOnDa LOD platform.

II. METHODOLOGY TO PUBLISH LINKED OPEN DATA

The subject of Linked Open Data publishing has been widely discussed in literature (e.g., [7]) and different projects and platforms have been developed to support this process. One of the first significant projects is Lucero and the resulting Tabloid toolkit, which aims to help institutions and developers to publish and consume linked data [8]. Another interesting work, supporting US open government data production and consumption, is the TWC LOGD portal [9]. A workflow for linked open data deployment is defined, consisting of different stages, where the conversion process is automated by using the csv2rdf4lod tool. A more recent initiative is represented by the Italian cultural heritage platform "dati.beniculturali.it", promoted by the Italian Ministry of Culture, which collects and publish standardized and interoperable LOD heterogenous datasets [10].

From a methodological point of view, a number of best practices, recommendations and guidelines have been produced. For example, Bauer and Kaltenböck [11] provide a step-by-step model, highlighting the most important issues that need to be taken into account in LOD publishing; W3C [12] presents best practices designed to facilitate LOD development and delivery; the "Agenzia per l'Italia Digitale" [13] proposes a general methodological approach for the interoperable opening of public data through the LODs. This methodology basically consists of the following steps: selection of dataset, data cleaning, analysis and RDF modelling, enrichment, interlinking, validation and publication.

The approach adopted in GIOCOnDa is in line with the above recommendations and in particular with the AGID guidelines. The selection of datasets was made on the basis of the results of a previous need analysis phase carried out in the project; as a starting point, data about museums, accommodation facilities and environment of the Insubric region are selected.

Concerning data cleaning, it is assumed that the selected datasets are already published as "clean" open data, where a quality check is already accomplished.

Once selected, datasets are deeply analyzed to understand their structure and appropriate ontologies and vocabularies are identified to model them.

In particular, the adopted ontologies are taken from the OntoPia network [14], also presented in [15]. They include for instance the Cultural-ON ontology for museums and the ACCO ontology for accommodations. In the GIOCOnDa LOD platform, data are imported from different sources and converted into the RDF format, according to these standard ontologies. The conversion process is detailed in the next section.

As additional steps, datasets are enriched with metadata and interlinked to other datasets. Metadata are added to the single datasets following the DCAT-AP standard. Interlinks are created to other datasets by identifying alignments and similarities between different datasets. For instance, a museum of the "Regione Lombardia" dataset can be declared "the same as" a museum described in Wikidata. The identification of interlinks is mainly carried out using the Silk software libraries [16].

Finally, datasets are published using Openlink Virtuoso Universal Server [17]; they can be queried through a SPARQL endpoint.

III. THE CONVERSION PROCESS FROM OPEN DATA TO LOD IN THE GIOCONDA PLATFORM

The core of the system lies in the mapping functionality of heterogenous data into linked open data, according to standard ontologies.

This conversion is a complex process that depends on the initial format and on the final standard RDF format. From a literature study it emerges that the most frequently adopted approach is the implementation of ad-hoc middleware. For example, to convert a relational database to LOD, a typical solution is to use declarative languages, such as D2R [18] or R2RML [19] that require ontological and programming skills.

In the GIOCOnDa LOD platform, the complexity of the conversion process is simplified by defining a converter, facilitated by a graphical user interface that an expert can use to configure the conversion. This process can be explained through a simple example: we would like to convert two different datasets about museums into a common interoperable format. The first dataset concerns *Lombard museums* retrieved from the Regione Lombardia portal in JSON format by means of REST APIs [20]. The second is represented by *Tessin Canton museums* retrieved from Wikidata through SPARQL queries.

Figure 1 shows an excerpt of the Lombard museums visualized on the Regione Lombardia portal, while Figure 2 shows an example of a Swiss museum in Wikidata [21].

To be able to configure the mapping from the original to LOD format, the structure of the two museum data sources has to be analyzed and an appropriate ontology selected. In this phase it is important to find the most appropriate ontology to model the domain. The Cultural-ON ontology [22] and its connected ontologies have been chosen because they are representative of the museum domain and can be exploited to support transnational interoperability.

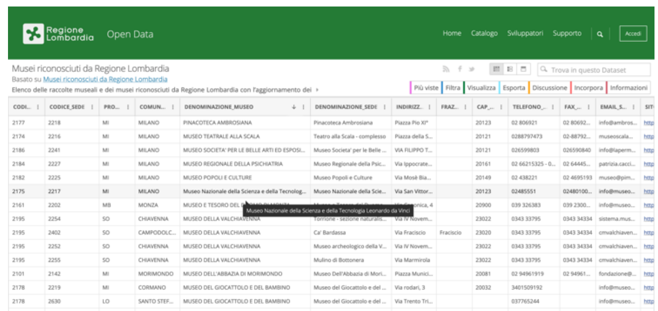

Figure 1. Lombard museums from the Regione Lombardia Open Data portal

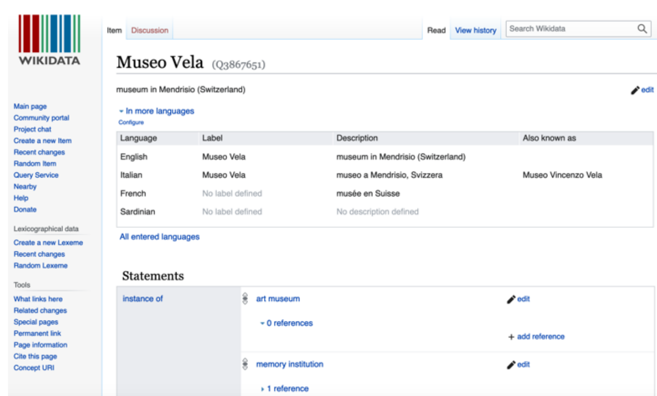

Figure 2. The Swiss *Vela* museum in Wikidata

The next step consists of going through the different descriptive fields of the museum datasets: for instance, each Lombard museum is described in terms of 79 fields, such as *denominazione museo* (name), *telefono* (telephone), *codice sede* (site code) as shown in Figure 1.

For each field, the objective is to find a match with the ontology classes and properties. For example, each museum could be represented as an instance of the c*is:Museum* class of the Cultural-ON ontology, where *cis* is the prefix of the

 ISBN: 978-1-61208-842-6

ontology namespace; the *telephone* field can be mapped into a property of a *smapit:OnlineContactPoint* instance of the Social Media / Contact and Internet ontology [23].

Figures 3 and 4 represent some result details of the conversion of a Lombard and a Tessin canton museum, respectively, in RDF Turtle, according to the Cultural-ON and the connected ontologies. In particular, in the excerpts, light blue highlights the *hasSite* relation to the *Site* instance, and yellow highlights the *hasOnlineContactPoint* relation to the *ContactPoint* instance with their respective properties.

It is important to note that two museums, initially available in different formats, are finally described in a common interoperable RDF format. This translation process leads, in this case, to information loss because there is not a full correspondence between the initial format and the ontological one. The ontology contains more classes and properties than the original file format; on the other hand, it is not enough expressive to represent all fields of the original data sources. For instance, the *number of visitors* is not included in the Cultural-ON ontology.

```
museum:Museo_2175_sede_2217_Museo_Nazionale_della_Scienza_e_della_Tecnologia_Leonardo_da_Vinci  a
cis:CulturalInstituteOrSite, cis:Museum ;
         rdfs:label "Museo Nazionale della Scienza e della Tecnologia Leonardo da Vinci" ;
         cis:institutionalName "Museo Nazionale della Scienza e della Tecnologia Leonardo da Vinci" ;
         cis:hasSite site:Sede_2217;

         smapit:hasOnlineContactPoint contactPoint:Contatti_Museo_Leonardo_da_Vinci ;

...
site:Sede_2217
         a cis:Site, poiapit:PointOfInterest;
         rdfs:label "Museo Nazionale della Scienza e della Tecnologia Leonardo da Vinci";
         cis:siteAddress address:Indirizzo_della_Sede_Museo_scienza_Leonardo_da_Vinci;
         clvapit:hasGeometry geometry:geometry_Museo_Leonardo_da_Vinci .
...

contactPoint:Contatti_Museo_Leonardo_da_Vinci
      a smapit:OnlineContactPoint ;
      smapit:hasEmail email:email_museo_Leonardo_da_Vinci;
      smapit:hasTelephoneNumber phone:phone_museo_Leonardo_da_Vinci ;
      smapit:hasTelephoneNumber fax:fax_museo_Leonardo_da_Vinci ;
      smapit:hasWebSite website:web_museo_Leonardo_da_Vinci .
```

Figure 3. Excerpt of the Lombard Museum of Science and Technology in RDF Turtle

```
museum:Museo_Q3867651_Museo_Vela   # Q3867651 ?subject usato per nome individuo insieme a label
        a cis:CulturalInstituteOrSite, cis:Museum ;
        rdfs:label "Museo Vela" ;
        cis:institutionalName "Museo Vela" ;
        cis:hasSite site:Sede_Q3867651 ;

        smapit:hasOnlineContactPoint contactPoint:Contatti_Museo_Vela .


site:Sede_Q3867651
        a cis:Site, poiapit:PointOfInterest ;
        rdfs:label "Museo Vela ";
        cis:siteAddress address:Indirizzo_della_Sede_Museo_Vela ;
        clvapit:hasGeometry geometry:geometry_Museo_Vela .
...

contactPoint:Contatti_Museo_Vela
      a smapit:OnlineContactPoint ;
      smapit:hasEmail email:email_museo_Vela;
      smapit:hasTelephoneNumber phone:phone_museo_Vela ;
      smapit:hasWebSite website:web_museo_Vela .
```

Figure 4. Excerpt of The Swiss Vela museum in RFD Turtle

In the conversion process, the mapping from the initial input data format to the final LOD format would need to be configured for each data source. This requires a deep knowledge of the OWL syntax, and understanding of the classes and datatype/object properties of the selected ontologies.

To simplify the conversion process, an internal vocabulary was created, with the objective to describe in a homogeneous and simple way data coming from different sources, without knowing the details of the ontology and further separate the input from the output. The main advantage of having this vocabulary is to hide the complexity of the ontology in the mapping management. The internal vocabulary is organized in categories, that represent contexts or ontologies; each category contains classes; each class has a number of fields. For instance, to describe museums we have defined the *museum Cultural-ON* category; this category contains classes, such as *museum* and *discipline*, and fields, such as *geographical coordinates*.

Thanks to the internal vocabulary, the conversion process is divided in two steps:

- the conversion from the input data format to the internal vocabulary (*input mapping*)
- the conversion from the internal vocabulary to the ontological LOD format (*output mapping*).

Going back to the museum example, the two datasets, originally described in different formats and with different descriptive fields, are translated by means of the input mapping specifications into a common format, which is described by the internal vocabulary. The resulting datasets are then converted to the LOD format, according to a standard ontology, by means of the output mapping specification. This guarantees standardization and semantic interoperability. Figure 5 illustrates the process.

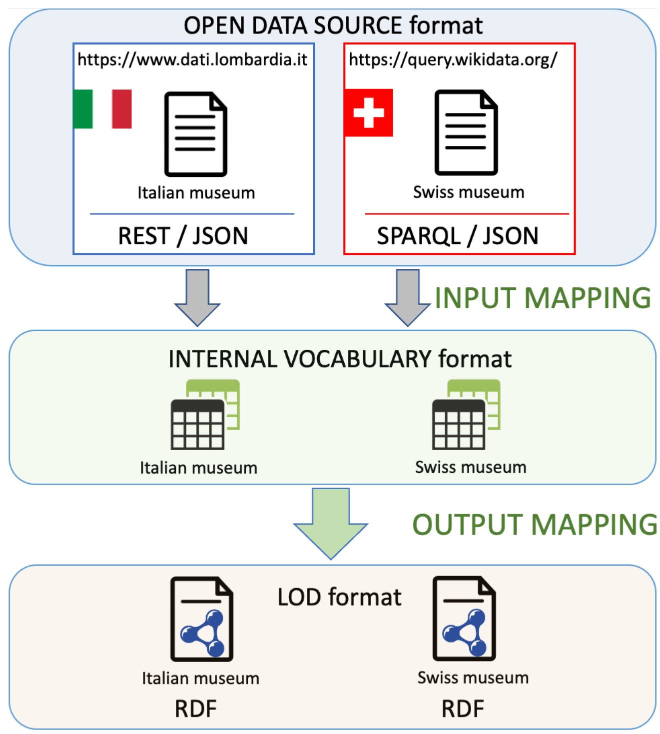

Figure 5. Two step conversion process: museum data example

While it is necessary to configure the input mapping of each imported dataset towards the internal vocabulary, the output mapping of a specific category (e.g., museum) to the corresponding LOD format has to be configured only once. The first step can be accomplished by a user who knows the input format, the domain and the internal vocabulary, the second step requires a deep knowledge of the ontologies and of the OWL language.

This mechanism, to convert Open Data to Linked Open Data based on two independent and configurable steps, is the peculiar feature of GIOCOnDa LOD platform compared to other frameworks, which do not provide the flexibility and dynamic configurability required by the GIOCOnDa project.

IV. GIOCONDA LOD PLATFORM

The GIOCOnDa LOD platform [24], implemented as a Java based web application, provides different functionalities that enable the publication of LOD datasets starting from open datasets, and their visualization in a catalogue or in a map.

The web app presents a menu consisting of different items: dataset catalogue, input mapping and output mapping.

A. LOD Datasets

The catalogue shows the list of the existing datasets, as shown in Figure 6, and enables the creation of a new LOD dataset by converting an existing open dataset on the basis of the input and output mapping configuration. The system supports dataset updates at regular intervals (e.g., for air quality measurements) and propagates the changes to the RDF representation.

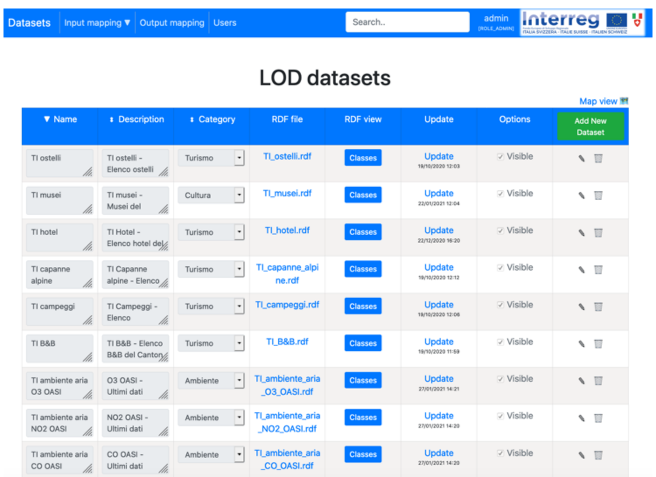

Figure 6. LOD datasets catalogue

By clicking on the "map view" button it is possible to visualize geo-locable data on the map as shown in Figure 7.

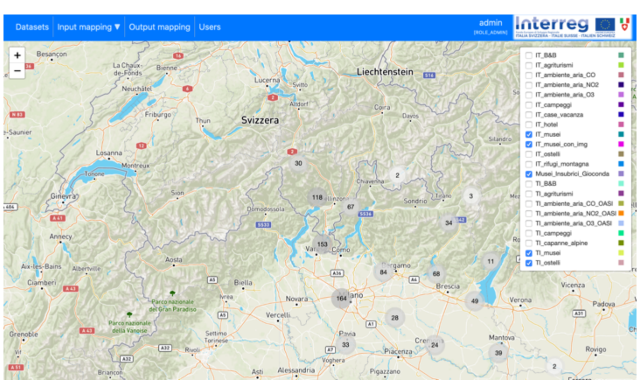

Figure 7. LOD Datasets visualized in the map

About 25 datasets about accommodation, museums and air quality have been boosted to LOD datasets through the GIOCOnDa platform, resulting as published LOD resources complaint to the selected ontologies and vocabularies. Concerning validation, the output mapping process guarantees by design and implementation that the produced datasets are serialized in RDF format conforming to each ontology. A further manual checking has been accomplished on some resources of each typology.

B. Input Mapping

The Input mapping concerns the configuration of the conversion from the input format to the internal vocabulary. Together with the output mapping it enables to configure the conversion from different input data sources to the final LOD format.

In general, the system accepts input data retrieved from a number of sources in different formats, such as JSON, CSV and XML, and using different services, such as Rest APIs, SOAP APIs and SPARQL Queries. For each input format and service the conversion towards the internal vocabulary is configured through the input mapping.

Figure 8 clarifies how the mapping mechanism works: the first column shows the fields of the original data source, the second and third columns concern the internal vocabulary, in particular the second identifies the category and the third the fields.

Categories and fields of the internal vocabulary are predetermined and selected from a drop-down menu, while the input fields must be entered by hand, according to the data structure in use.

In the example shown in Figure 8 we work with fields of the "Cultural-ON museums" category; the "email_sede" field of the initial source, representing the email of the museum site, for example, is mapped into the "email" field of the "museum Cultural-ON" category of the internal vocabulary. In some cases, it is possible to group multiple fields of the data source into a single field of the intermediate vocabulary; for example, to compose the field "full_address", more fields of the input source are used.

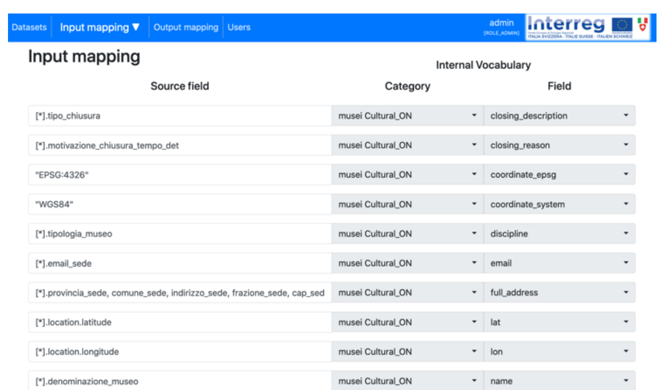

Figure 8. Input mapping

In this mapping, particular attention is dedicated to how the geospatial data are represented. This is essential to guarantee interoperability and efficient sharing of information across different regions and national standards. The Cultural-ON ontology assumes, by default, that spatial data are represented in the geocentric Datum WGS84 and that the coordinates are expressed in terms of latitude and longitude. Therefore, data are transformed in this system when they are imported in the GIOCOnDA platform and appropriate metadata are added to make the Coordinate Reference System (CRS) explicit.

C. Ouput Mapping

From the output mapping page, it is possible to create, modify and extend the internal vocabulary and define its mapping to the ontology. This process requires a deep knowledge of ontological concepts and existing reference ontologies. Nevertheless, this mapping has to be done only once for each category by an expert.

As already said, the internal vocabulary consists of several categories, similar to contexts or ontologies; each category contains classes, with associated a number of fields. Examples of categories include museums, addresses, accommodations, etc.

As shown in Figure 9, the output mapping defines the match between internal vocabulary classes and ontology classes, and between fields of the internal vocabulary and object and datatype properties of the ontology; this is visible by activating the "Show fields" button.

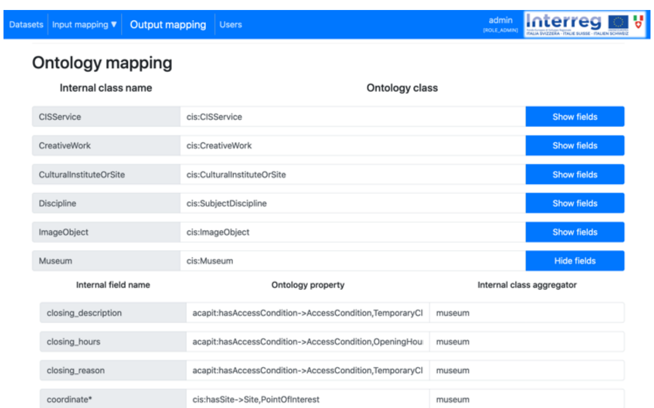

Figure 9. Output mapping: class/field match

It is worth noting that only categories and fields of the intermediate vocabulary defined in the output mapping can be used in the input mapping (but not classes), providing in this way a simplified version of the data structure for the non-expert user.

A specific interlinking software has also been developed and integrated in the GIOCOnDa LOD platform. In order to boost a dataset to 5* level it is necessary to configure the interlinking and then activate it. The configuration is accomplished in the output mapping page where one or more interlinking files can be associated to each category. This file is generated using the SILK Link Specification Language and contains rules to create cross-reference links towards external datasets, such as Wikidata. The activation takes place in the datasets page, where it is possible to enable or disable interlinking on a specific dataset.

V. Conclusions and Future Works

This paper has presented a platform that facilitates the process of conversion of open data to linked open data, by means of a visual interface, without the need of a specific software programming. Indeed, one of its main advantages is the reduction of complexity of a process that requires deep knowledge of ontologies and programming skills. The complexity of mapping existing data to standard ontologies is one of the major issues preventing a larger diffusion of LOD.

The GIOCOnDa LOD platform contains a number of conversion configurations that allow different data sources to be converted to LOD in different domains. For instance, if a dataset has the same structure of an existing one (for example, a new dataset structured as the Lombard museums), the conversion is very simple, since the input mapping is very similar to an existing one and the output mapping is already defined.

However, the platform is also flexible and extensible, and enables to import and convert other datasets: for example, the conversion to LOD of a new dataset about museums, with a structure that can be mappable to the existing internal vocabulary, only requires the configuration of the input mapping from the initial format to the internal vocabulary, because the output mapping is already configured. More labour-intensive but still possible is to convert a dataset with a new structure, not mappable on the existing internal vocabulary; for example, a new dataset about bike sharing. In this case, it is necessary to find an ontology that models the domain; to extend the internal vocabulary with a new category and define the mapping towards the ontology - this is configured in the output mapping; to define the mapping from the original data format to the internal vocabulary - this is configured in the input mapping.

In addition to the conversion to LOD, another important step of the adopted methodology is the identification and creation of interlinks between datasets. A specific interlinking software has been developed to configure and activate the process of identification of cross-reference links towards external datasets. The integration of the interlinking module in the GIOCOnDa LOD platform enables lifting datasets to 5* level, creating added value through an Extract-Transform-Load (ETL) pipeline. This has been demonstrated for instance in a showcase that presents data about museums of the Insubric region taken both from the GIOCOnDa LOD datasets and from Wikidata.

In spite of the benefits offered by platform to publish LOD datasets (visual interface, configurability, extensibility), it also presents some limitations: the main one is the possibility of information loss during the conversion to LOD if there are fields not represented in the selected ontology. A possible solution is the extension of the selected ontologies with additional fields and the publication of the new version with appropriate documentation.

Finally, the platform development is still in progress and some details need to be fully implemented or considered for future implementation, such as the validation process, inference and interlinking.

ACKNOWLEDGMENT

We acknowledge the European Regional Development Fund, the Tessin Canton and the Swiss Confederation for the financial support provided to the project. We also acknowledge all the partners involved in the project for their contribution to the different project activities.

REFERENCES

[1] GIOCOnDa, Integrated and holistic management of the Open Data life cycle (Gestione integrata e olistica del ciclo di vita degli Open Data) https://interreg-italiasvizzera.eu/progetti/gioconda/ [retrieved: March, 2021].

[2] Berners-Lee, Linked Data, http://www.w3.org/DesignIssues/LinkedData.html, 2009 [retrieved: March, 2021].

[3] Open Data, Regione Lombardia, https://dati.lombardia.it [retrieved: March, 2021].

[4] Arpa Lombardia, https://www.arpalombardia.it/ [retrieved: March, 2021].

[5] Ticino Turismo, https://www.ticino.ch [retrieved: March, 2021].

[6] OASI, Osservatorio Ambientale della Svizzera italiana, https://www.oasi.ti.ch [retrieved: March, 2021].

[7] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers, 2011.

[8] The Lucero Project, https://lucero-project.info/ [retrieved: March, 2021].

[9] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, J. A. Hendler, “TWC LOGD: A portal for linked open government data ecosystems”, Journal of Web Semantics, vol. 9(3), pp. 325-333, 2011, doi: 10.1016/j.websem.2011.06.002.

[10] Open Data e Linked Data, https://www.beniculturali.it/open-data-e-linked-data [retrieved: March, 2021].

[11] B. Florian and K. Martin, Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers. Edition mono/monochrom, Vienna, Austria, 2012, ISBN: 978-3-902796-05-9,

http://www.semantic-web.at/LOD-TheEssentials.pdf [retrieved: March, 2021].

[12] W3C, Best Practices for Publishing Linked Data, W3C Working Group Note 09 January 2014, https://www.w3.org/TR/ld-bp/ [retrieved: March, 2021].

[13] Agenzia per Italia Digitale, Guidelines for semantic interoperability through Linked Open Data (Linee Guida per Interoperabilità semantica attraverso i Linked Open Data), http://www.agid.gov.it/sites/default/files/documentazione_trasparenza/cdc-spc-gdl6-interoperabilitasemopendata_v2.0_0.pdf 2012 [retrieved: March, 2021].

[14] Ontologie e Vocabolari Controllati, https://github.com/italia/daf-ontologie-vocabolari-controllati [retrieved: March, 2021].

[15] G. Lodi, “OntoPiA –The network of ontologies and controlled vocabularies for public admistration (OntoPIA - La rete di ontologie e vocabolari controllati per la pubblica amministrazione)”, Open Data Sicilia – raduno annuale, 9/10 novembre 2018, http://ods2018.opendatasicilia.it/presentazioni/Lodi-OntoPiA.pdf, 2018 [retrieved: March, 2021].

[16] Silk, The Linked Data Integration Framework. http://silkframework.org/ [retrieved: March, 2021].

[17] Open Link Virtuoso, https://virtuoso.openlinksw.com [retrieved: March, 2021].

[18] C. Bizer and R. Cyganiak, “D2R Server - Publishing Relational Databases on the Semantic Web”, 5th International Semantic Web Conference (ISWC 2006), Athens, USA, November 2006, http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf, 2006 [retrieved: March, 2021].

[19] R2ML: RDB to RDF Mapping Language https://www.w3.org/TR/r2rml [retrieved: March, 2021].

[20] Musei riconosciuti da Regione Lombardia https://www.dati.lombardia.it/Cultura/Musei-riconosciuti-da-Regione-Lombardia/3syc-54zf [retrieved: March, 2021].

[21] Museo Vela https://www.wikidata.org/wiki/Q3867651 [retrieved: March, 2021].

[22] Cultural-ON (Cultural ONtology): Cultural Institute/Site and Cultural Event Ontology, https://w3id.org/italia/onto/Cultural-ON [retrieved: March, 2021].

[23] Social Media / Contact and Internet ontology - Italian Application Profile https://w3id.org/italia/onto/SM [retrieved: March, 2021].

[24] GIOCOnDa LOD platform https://gioconda.supsi.ch/ [retrieved: March, 2021].