CROSS-BORDER OPEN DATA SHARING: GIOCOnDA PROJECT

J. F. Toro [1, *], D. Carrion [1], A. Albertella [1], M. A. Brovelli [1],

[1] Politecnico di Milano, Civil and Environmental Engineering Department – Section of Geodesy and Geomatics, 20133 Milan, Italy
– juanfernando.toro@mail.polimi.it, (daniela.carrion, alberta.albertella, maria.brovelli)@polimi.it

Commission IV, WG IV/4

KEY WORDS: Interreg, data interoperability, Open Data, DCAT, GeoDCAT, Semantic Web

ABSTRACT:

Open Data, and Open Government Data, are proving to be an important resource for the economic development inside the domain where information has a key role (Carrara et al., 2015). Although, different practices for data publishing have led to misalignment, underuse and repetition of information (Bizer et al., 2011). For this reason, the Public Administrations have undergone efforts on integrating the information and promoting interoperability through the implementation of best practices, as for example, the use of a common semantics vocabulary for the metadata (DCAT) as proposed by the ISA2 programme of the European Commission. The Interreg Italy-Switzerland GIOCOnDA project has been proposed for enhancing the data sharing processes in the cross-border area, particularly addressing tourism and mobility that are key economic activities for the region. For this work, a review on the data catalogues published in dati.lombardia.it and opendata.swiss is presented. The revision of the datasets showed the need for: 1) defining common semantics for the description of the categories of data to avoid the arbitrary use of vocabularies, and 2) adopting standards for the description of geodata. On the other hand, it was observed the potential to gather existing information to produce geodata querying the datasets with specific keywords that can provide spatial information. Open data, as well as the use of best practices for publishing data, push towards the use of FOSS. In this work, Python has been exploited to analyse the content of the catalogues to access web portals resources.

1. INTRODUCTION

Open data, and, specifically, open government data, have become a mean for the Public Administrations (PAs) to prove transparency on their activities through the engagement of the public in their work. In addition, the PAs have also recognised that dissemination of socio-economic data to the public have an impact on the business and citizens' economic activities under their governance. For instance, it was estimated that by year 2020 the direct market value generated from the re-use of the available data could be around 76 billion euros, and, from this amount, 22 billion could correspond to the PAs (Carrara et al.,2015). However, it was clearly mentioned that this would be achieved through the commitment of the PAs with respect to the challenges that can arise from an efficient and unique data publishing for the reuse of data. Then, one of the tasks to follow is the creation of a framework which can guide to ease the exchange, and that can provide expectations, context and meaning of information that can be found on the resources. Then, the data publication issue is now the data interoperability.

The European Union has been actively developing different programmes to promote data interoperability among the PAs. Initially, the European Commission presented the Interoperability solutions for public administrations programme (ISA, 2010-2015) "…to improve the interactions, exchange and cooperation among European public administrations" (European Commission, 2010). Later, it has been adopted the continuation of the programme Interoperability solutions for public administrations, businesses and citizens (ISA2, 2016-2020), to foster the reuse of information by the public, the data publication processes and the provision of additional

information in existing datasets. Furthermore, these programmes have developed initiatives for the data interoperability, such as the European Interoperability Strategy (EIS) and the European Interoperability Framework (EIF), establishing the need for a standardised semantic vocabulary (European Commission, 2010; European Commission, 2019). On 2016, ISA2 through the EIF, presented the action "SEMIC: Promoting semantic interoperability among the EU Member States" for the use of semantic tools and introducing the Data Catalogue (DCAT) semantic vocabulary defined by the World Wide Web Consortium (W3C[(1)]; European Commission, 2016b).

It is important to realise that the current actions to improve the practices for exchange of open government data are performed at larger scale level than the national administrations, addressing the possibility of knowledge exchange between the cross-border regions. This is due to the awareness of the PAs of the fact that some of the production activities require the interaction with stakeholders of different governances. Then, the datasets integration, not only fosters the reuse and improves the efficiency, it should also reduce the asymmetries of information availability between the cross-border regions.

Over the past years, the cooperation between Italy and Switzerland has led to the implementation of several Interreg projects aiming towards the exchange of information in agreement with the programmes developed by the European Commission.

---

* Corresponding author
[(1)] https://www.w3.org/TR/vocab-dcat/

Here, it is presented the ongoing review of the current state of the data infrastructure and open data availability between the Italy-Switzerland cross-border regions, with particular focus on data interoperability.

2. GIOCONDA PROJECT

The GIOCONDA (*Gestione Integrata e Olistica del Ciclo di vita degli Open Data*, Integrated and holistic management of Open Data life cycle) Interreg Italy-Switzerland project surges as a proposal for compiling the open data sources within the Insubria Region under a single infrastructure and, with this, to stimulate the territory economy. The territory of interest comprehends the cross-border region between Italy and Switzerland, including Lombardy Region and Canton of Ticino. This territory relies on tourism as one of the main economic sources, for this reason, the study focus is based on it.

In the area of interest, several of the productive activities have been focused on the tertiary sector of production, that corresponds to the offer of services and markets of goods (ARE, 2011). Under those circumstances, there has been a job concentration on the urban centres, along the alpine area, with an estimate of 71% (ARE, 2011). In 2011, it was estimated that the jobs of the tertiary sector with respect to the total is between 50%-75%, where, the highest values correspond to the major cities involved in touristic activities (Working Group Demography and Employment of the Alpine convention, 2011). A key point for the economic development of the cross-border area is the appropriate supply of information on people mobility to support the decision-making activities, considering that touristic activities are often seasonally driven.

Tourism industry is characterised by a rapid developing environment, reflected in needs for fast adaption to sustain the economic activities of businesses, on which particular attention has been devoted (European Commission 2010a; European Commission 2012; European Commission 2016a, Vecchio et al., 2018). These activities are dominated by the knowledge on best practices and information dissemination (Makkonen et al., 2018). For this reason, the provision of information to everyone is important to foster synergies and to permit the economic growth of the area.

In order to mitigate misalignment and underuse of common information in cross-border regions and to reduce the burden of maintaining databases in an open and shared standard format, the two cross-border PAs must cooperate. GIOCOnDA project proposes the construction of a common information framework that starts from the data held by the PAs and the territory of the Insubric Region, based on a same semantic infrastructure following the guidelines exposed by the ISA2.

The PAs of the two cross-border territories have been asked to work together in order to adopt a model that will allow the production and exposure of data according to standards to foster interoperability, in a common information framework, useful to satisfy the information needs of the Insubric Region.

This document focuses on two of the most relevant Open Government Data sources of the Insubria Region: Lombardy Region[2] and Swiss Federal Statistics Office[3]. Yet, it should be

---

[2] https://dati.lombardia.it/
[3] https://opendata.swiss/it/

remarked that the GIOCOnDA project will work to involve additional open data providers.

3. DATA ANALYSIS

This section describes the methodology followed to analyze the data catalogues with the purpose of identifying the datasets that can be translated into geodata. Firstly, by reviewing the syntax and the semantics of the metadata in the catalogues in order to understand which kind of information is available for the datasets. Secondly, by performing a comparison between the data catalogues for the specific vocabulary used to describe the datasets. At last, it is extracted detailed information of the datasets related to position from which the geolocation of data can be derived.

The datasets information, for both catalogues, was retrieved with the extensive use of Python and readily implemented APIs on the web portals. In a similar manner, through the use of Python, it was possible to access the metadata through the identifiers of each dataset, and to later retrieve content to figure out the availability of spatial information.

3.1 Metadata Semantic Vocabulary – DCAT-AP

In literature, different issues regarding interoperability have been identified: process, legal technical, syntax, semantic and query (European Commission, 2010b). For instance, here syntax and semantic issues are discussed, corresponding to data structure and data description.

Initially, the implemented syntax and semantics for the published datasets on each data catalogue have been assessed. The participation on the ISA2 program encourages the use of a common set of practices for the description of the metadata (DCAT-AP, Data Catalogue Application Profile, that is the interchange format for metadata in the European Union; Perego, et al., 2016). Due to this, there is a recognized structure in which information should be provided for both catalogues, requiring the compilation of at least some mandatory fields. Eventually, thanks to the flexibility of the DCAT-AP, each of the countries involved in the program can set up their own DCAT-AP according to the information needs. For instance, there are the DCAT-AP-IT (AgID, 2017) and the DCAT-AP-CH (OGD-CH, 2016), that are the data catalogue application profiles for Italy and Switzerland, respectively.

The advantages of working on a unique standard is reflected on the reduction of efforts to identify information available in the datasets. For instance, the representation of two datasets with different semantics require the need to develop a common framework for their description, this process is referred to as harmonization (Veeckman et al., 2017). This is avoided thanks of the efforts of the European PAs.

Currently, the DCAT-APs for each of the countries request for similar information for the datasets such as identifier, description, last update, theme, publisher, frequency of updates, download URL, etc. However, the presence of geospatial data is not accounted as mandatory information to be provided, but as optional inputs. From this, two main problems may arise: 1) some of the submitted datasets containing spatial information are published as texts formats with coordinates attributes and 2) the semantics used by each of the DCAT-Aps can be different. For now, the aim of this analysis is checking for the possibility to convert data to geodata. Regarding the semantics

misalignment, Bizer et al. (2011) have mentioned the issues that can carry the integration of formats to already existing semantic structures, for this reason, the creation of a complementary semantic structure was needed (GeoDCAT-AP) at European level. The conversion from data to geodata must ensure the interoperability, thanks to the compliance to specific semantics, according to the GeoDCAT-AP guidelines. The Italian national data[4] and geodata[5] web portals by *Repertorio Nazionale dei Dati Territoriali* (RNDT, AgID, 2019) provide an example of interoperability effort to which GIOCOnDA project may refer.

3.2 Data catalogues access – SODA & CKAN

At present, the adoption of semantic standards for the metadata representation has eased the delivery of APIs for publishing, querying and updating datasets. In this case, the data catalogues management are done through Socrata Open Data API (SODA), for Lombardy Region, and CKAN, for the Swiss Federal Statistics Office portal. These two APIs can give access to the same information but are constructed in a different way and access the data differently. SODA is constructed as a REST (Representational State Transfer) API, differently, CKAN API is constructed as a RPC (Remote Control Procedure).

To mention the positive, the existence of the APIs allows to access both, data and metadata, but, as previously mentioned, the data acquisition should be performed in a different way. In the case for the REST API the data acquisition is quite straightforward as the resources, for example metadata, can be accessed by means of the identifiers of the datasets. While, the RPC APIs require, often, to perform multiple queries, which are predefined and already implemented in the server, in order to retrieve the desired content.

This section aims 1) at identifying the vocabulary employed for the description of the datasets and 2) at understanding which of the datasets can be potentially transformed into geodata. These specific tasks were carried out with the use of the FOSS Python to retrieve specific data of interest, to do so, it has been necessary to prepare the queries according to the specific API structure. The formats in which the datasets are commonly accessed on these catalogues are tabular, CSV, or object-oriented formats, JSON or GeoJSON. The analysis carried out used the functions for reading JSON format resources.

3.3 Data catalogues comparison

A comparison has been performed between the Italian and Swiss datasets to assess the coherence of the catalogues' categories.

3.3.1 Datasets themes: In general, the DCAT-AP-IT and DCAT-AP-CH request for the same format type to describe each of the mandatory fields. Nevertheless, the semantics used to describe the datasets is somewhat different from one to another making more complex the treatment of the datasets.
For example, the themes used for categorizing the datasets are 23 for Lombardy Region and 24 for the Swiss Federal Statistics Office. The difference is not only in the number of themes, but, more importantly, there is a misalignment in the vocabulary used to define them (see Figure 1 where possible correspondences have been found among the two data structures).

Figure 1. Misaligned categories, under the dct:theme component for the dcat:dataset class for the Lombardy Region and Swiss data catalogues

Then, there is need to stablish a well-defined vocabulary to align the description of the datasets. It must be remarked that the equivalency presented in the figure, is an approximation based on the similarity of the terms. Yet, in some cases, the datasets assigned to a category in one catalogue can correspond to a different one in the other catalogue. For this reason, the description of the datasets should be revised to find a common ground.

3.3.2 "Tourism" and "Mobility and transport" – Keywords: Due to the key interest of tourism, mobility and transport for the economic development, the corresponding datasets have been explored.

For now, the comparison between the data catalogues was made based on Italian language which is shared between the two data providers (the swiss catalogue provides four different languages German, English, French and Italian for the description of the datasets). The available datasets under the analysed categories are summarised in Figure 2.
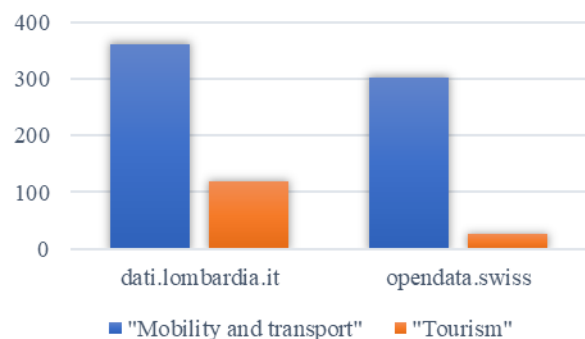

Figure 2. Count of datasets under the "Tourism" and "Mobility and transport" categories

A first approach for the comparison is to inspect which of the keywords on each of the datasets are most commonly used for the description of the content. This is done to have an idea on which are the PAs expectations with respect to the user needs and to check if these are equivalent from Italian and Swiss side.

Table 1 summarizes the most frequently used keywords under the "tourism" category. One thing that can be observed from this preliminary analysis is that the focus of Lombardy Region Web portal regarding tourism is more devoted to services while, in contrast, the Swiss Web portal refers more the fluxes of people inside their territory. A second point to highlight, and which is not observable in the table, is that the Swiss data catalogue presents only 11 tags in Italian, but in German language 52 different tags are available. So, it is evident the need to perform additional equivalency of semantics between different languages.

dati.lombardia.it		opendata.swiss	
Turismo (Tourism)	91	*tabella* (table)	10
Alberghi (Hotels)	75	*turismo* (tourism)	10
Allogi (Lodgins)	72	*rilevazione-totale* (total survey)	6
Camera (Room)	72	*statistica-della-ricettivita-turistica* (statistics of tourist accomodation)	6
Commercio (Commerce)	72	glam	5
o1	70	openglam	5
giovani (youth)	8	*statistica-di-elezioni-e-votazioni* (statistics of elections and votes)	4
itinerari (itinerary)	8	*micro-censimento.mobilita-e-trasporti* (mobility and transport micro-census)	1
low cost	8	*mobilita* (mobility)	1
n1	8	*rurale* (rural)	1
ostelli (hostels)	8	*tipologia-dei-comuni* (types of municipalities)	1

Table 1 "Tourism" theme, top 10 count, most-used keywords in the data catalogues

Similarly, the frequency analysis for the usage of keywords under the "mobility and transport" category has been performed (see Table 2). An interesting point is that the vocabulary used to describe the "mobility and transport" by Lombardy Region open data catalogue is similar to the one used by the Swiss PA to address "tourism". In fact, the approach to "mobility and transport" by Lombardy Region addresses mainly the fluxes of people, like "tourism" for Swiss data. With this in mind, it is important to account the possibility to revise the existing guidelines for the description of the datasets so that they can be classified accordingly on a unified ontology.

dati.lombardia.it		opendata.swiss	
Mobilità (Mobility)	131	*reti-di-trasporto* (transport network)	20
Trasporti (Transports)	127	*pianificazione-della-conservazione-e-dellarchiviazione-aap--confederazione* (conservation planning and archiving-aap ---	11

		confederation)	
s1	106	*geodati-di-base* (basic geodata)	10
destinazione (destination)	102	*strada* (street)	10
Matrice (matrix)	102	*trasporto-pubblico* (public transport)	10
origine (origin)	102	glam	9
spostamenti (movements)	102	openglam	9
statistica (statistics)	102	*trasporti* (transports)	9
Parcheggi (Parkings)	31	*geoportale-e-geoch* (geoportal and geoch)	8
sosta (stop)	19	*Ferrovia* (Railway)	7

Table 2. "Mobility and transport" theme, top 10 count, most-used keywords in the data catalogues.

3.3.3 Datasets identification for conversion to geodata: To be able to understand which datasets can give additional information that can potentially be translated into geodata, a query of specific terms related to the description of positioning has been performed. It was found that some of the observations within the datasets under "tourism" contained the attribute fields presented in Table 3.

Field	Description	Format
"indirizzo" ("Address")	Address of the observation. String.	String
"location"	Couple of latitude and longitude.	List of floating-point numbers
"lat", "long"	Latitude and longitude.	Floating-point numbers
"geo_x", "geo_y"	Coordinates x and y.	Floating-point numbers

Table 3. Fields within the datasets of "tourism" associated to position information

With the aid of Python, it was possible to identify the datasets which contain the fields in Table 3, as presented in Figure 3.
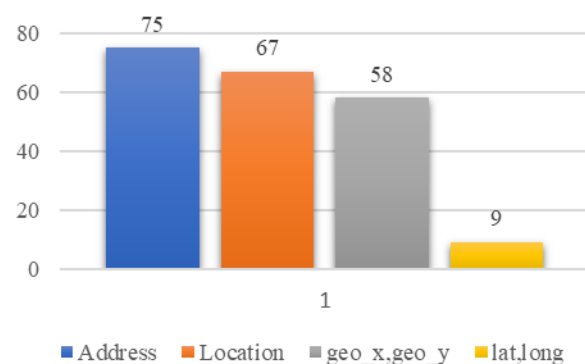

Figure 3. Datasets under "tourism" category that include position attributes.

All datasets with positioning information provide information about the "Address". Then, it was found that 75 out of the 118 datasets belonging to "tourism" category have information that can contribute to derive geodata. In particular, the datasets which information about the location correspond to hosting services like hotels, hostels, lodgings, etc. In addition to these

data in tabular format, which contain geolocation information, Regione Lombardia provides as well geospatial data in geospatial format for download or through a Web Service.

On the other hand, a similar analysis for the Swiss datasets was not possible for the tabular data of "tourism" category because information is provided about population mobility rather than specific location of services. Although, the Swiss catalogue allows the download of datasets in geospatial format or their visualization through Web Map Services (WMS). For this, the identification of the datasets which can be potentially used for their conversion into geodata has to be continued, considering also the possibility to collect relevant information for tourism from data classified under different themes.

4. DISCUSSION

This document presents first analyses made for the Interreg Italy-Switzerland project GIOCOnDA, being the project main purpose the integration of the open data infrastructure of the Public Administrations along the cross-border area. The knowledge flows across the border can provide expected benefits to the productive activities in the territory, with the possibility of adding an extra value making available, when possible, the geospatial component to existing datasets. For the time being, the following to authoritative sources have been taken into account: dati.lombardia.it and opendata.swiss. Three main challenges were identified. 1) The need to elaborate the metadata semantics vocabulary in such a way that the data are appropriately handled, 2) the definition of shared vocabulary for the description of the datasets and 3) the possibility to transform existing datasets into geodata.

Initially, the analysis identified the semantic interoperability between the metadata description for both open data catalogues. This has been possible thanks to the initiatives proposed within the European Union for the data interoperability, namely, the adoption of the DCAT-AP vocabulary for metadata standards. Additionally, each country has proposed additional, optional, attributes that can describe the application profiles including geospatial features to comply with the publishing of geodata. However, in most of the cases the datasets are supplied to the Web portal in table formats containing position attributes and not in geospatial format. The first challenge on achieving interoperability is the need to adopt standardised semantics that support geospatial formats, which is an ongoing activity for the European Commission that is the GeoDCAT.

For the time being, the semantics employed in the datasets under study were not homogenous and the keywords definition change between catalogues. Then, the second challenge for achieving interoperability is a common definition of the semantics to describe the datasets.
Nevertheless, it is important to remark the possibility to address the existing keyword dictionaries to identify the information that the Public Administrations expect to be requested.

Regarding the possibility to convert, when possible, tabular data into geodata, it has been verified that, in Regione Lombardia dataset, around 63% of the data under the "tourism" category already provide information about the position. Both Regione Lombardia and opendata.swiss provide also data which are already in geospatial format. The third challenge consists in the possibility of providing geodata, under the same standards for the description of the metadata to provide a uniform data

semantics that facilitates the data search of the central Web portal for Insubria Region.

Another point worth to be mentioned, is the spatial scale in which the data are provided. Then, a next step is also to understand the requirements to select the appropriate spatial level of detail for providing the information, should it be given in a canton scale, province scale, city scale, neighbourhood scale, etc.

In addition, it will be possibile to introduce different open data sources, both authoritative and non-authoritative, which could provide additional information on the development of complementary geodata.

ACKNOWLEDGEMENTS

REFERENCES

Agenzia per l'Italia Digitale (AgID). (2017). Datipubblici - linee guida cataloghi dati. Retrieved from https://linee-guida-cataloghi-dati-profilo-dcat-ap-it.readthedocs.io/it/stable/index.html

Agenzia per l'Italia Digitale (AgID). (2019, February 26). Dati territoriali e dati aperti: Il nuovo endpoint GeoDCAT-AP del Repertorio Nazionale (RNDT). Retrieved from http://eventipa.formez.it/node/179239

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts (pp. 205-227). IGI Global.

Carrara, W., Chan, W. S., Fischer, S., & Steenbergen, E. V. (2015). Creating value through open data: Study on the impact of re-use of public data resources. European Commission.

European Commission. (2010a). Europe, the world's No. 1 tourist destination: A new political framework for tourism in Europe.

European Commission (2010b). The European Interoperability Framework with 4 levels: legal, organisational, semantic and technical.

European Commission. (2012). Delivering on a new European neighbourhood policy. Brussels: European Commission.

European Commission. (2016a). Sustainable transnational tourism products. Retrieved May 20, 2019, from http://ec.europa.eu/growth/sectors/tourism/offer/sustainable/transnational-products_en

European Commission (2016b). DCAT Application Profile for data portals in Europe. (2016). Retrieved from https://joinup.ec.europa.eu/release/dcat-ap/12

European Commission (2016c). GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in

Europe. Retrieved from https://joinup.ec.europa.eu/sites/default/files/distribution/2016-08/geodcat-ap_v1.0.1.pdf

European Commission. ISA2 Programme. SEMIC action. (2018). DCAT Application Profile for data portals in Europe v1.2. Retrieved from https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/releases

European Commission. Directorate-General for Informatics (DIGIT). (2019). Use cases and benefits of ISA² specifications - ec.europa.eu. Retrieved from https://ec.europa.eu/isa2/sites/isa/files/docs/publications/isa2_study_use_cases_and_benefits_of_isa2_specifications_january_2019.pdf

Janečka, K., Čerba, O., Jedlička, K., & Ježek, J. (2013). Towards Interoperability of Spatial Planning Data: 5-Steps Harmonization Framework. Informatics, Geoinformatics and Remote Sensin.

Makkonen, T., Williams, A. M., Weidenfeld, A., & Kaisto, V. (2018). Cross-border knowledge transfer and innovation in the European neighbourhood: Tourism cooperation at the Finnish-Russian border. Tourism Management, 68, 140-151. doi:10.1016/j.tourman.2018.03.008

Open Data Government Switzerland (OGD-CH). (2016). DCAT-AP for Switzerland format. Retrieved from https://handbook.opendata.swiss/en/library/ch-dcat-ap

Perego, A., Friis-Christensen, A., & Lutz, M., Smart Descriptions & Smarter Vocabularies (SDSVoc) (2016, November 23). GeoDCAT-AP: Use cases and open issues. Retrieved from https://ec.europa.eu/jrc/en/publication/geodcat-ap-use-cases-and-open-issues

Swiss Federal Office for Spatial Development (ARE). (2012). Sustainable Mountain Development in the Alps. Retrieved May 25, 2019, from http://www.fao.org/fileadmin/user_upload/mountain_partnership/docs/ALPS FINAL 20120228 RIO Alps.pdf

Vecchio, P. D., Mele, G., Ndou, V., & Secundo, G. (2018). Open Innovation and Social Big Data for Sustainability: Evidence from the Tourism Industry. *Sustainability,10*(9), 3215. doi:10.3390/su10093215

Veeckman, C., Jedlička, K., De Paepe, D., Kozhukh, D., Kafka, Š., Colpaert, P., & Čerba, O. (2017). Geodata interoperability and harmonization in transport: a case study of open transport net. Open Geospatial Data, Software and Standards, 2(1), 3.

Working Group Demography and Employment of the Alpine Convention (2011) Preliminary paper to the XI Alpine Conference.

Revised May 2019