Open Government Data: Fostering Innovation

Ivan Bedini
Trento Rise
i.bedini@trentorise.eu

Feroz Farazi
University of Trento
farazi@disi.unitn.it

Ivan Tankoyeu
University of Trento
tankoyeu@disi.unitn.it

David Leoni
University of Trento
david.leoni@unitn.it

Stefano Leucci
University of Trento
stefano.leucci@unitn.it

Francesca Gleria
Provincia Autonoma di Trento
francesca.gleria@provincia.tn.it

***Abstract*— The provision of public information contributes to the enrichment and enhancement of the data produced by the government as part of its activities, and the transformation of heterogeneous data into information and knowledge. This process of opening changes the operational mode of public administrations, leveraging the data management, encouraging savings and especially in promoting the development of services in subsidiary and collaborative form between public and private entities. The demand for new services also promotes renewed entrepreneurship centered on responding to new social and territorial needs through new technologies. In this sense we speak of Open Data as an enabling infrastructure for the development of innovation and as an instrument to the development and diffusion of Innovation and Communications Technology (ICT) in the public system as well as creating space for innovation for businesses, particularly SMEs, based on the exploitation of information assets of the territory. The Open Data Trentino Project has initiated and fosters the process of opening of public information and develops as a natural consequence of this process of openness, the creation of innovative services for and with the citizens. In this paper we present how our project acts on long-chain, from raw data till reusable meaningful and scalable knowledge base that leads to the production of data reuse through the implementation of services that will enhance and transform the data into information capable of responding to specific questions efficiency and innovation.**

Keywords: OGD, innovation, entity, open entity, open data, big data

I. INTRODUCTION

The process of opening the data in Public Administrations (PA) as process of enhancement of public information implies a radical change to the data approach and work inside the PA. This takes time, an improvement of the data management methodology, creation of operational tools and providing a reliable space for sharing. Governments of various countries and administrative divisions thereof worldwide have nowadays been starting releasing a huge quantity of datasets in the context of Open Government Data (OGD) movement [1]. Along this line, a large, diverse and interesting collection of datasets are already published by the Autonomous Province of Trento (PAT) as OGD. New datasets are slowly becoming available and the existing ones are updated whenever needed for the purposes such as correcting mistakes and adding new data horizontally (as instances) or vertically (as properties). The data catalog is also linked with the website of the Department of Innovation of the PA[1]. This department governs the process of opening new data and the dissemination of the so called data culture. This is an important result to reach both inside and outside of the PAT. People started understanding the value of publishing high-quality data and the power in the reuse of them. Linking data will highly foster the value of sharing data. It will also lead to a new kind of data-centric public bodies that will empower citizens and generate innovative services.

Immense numbers of government datasets could open up new opportunities for application developers and trigger game changing disruptive business models to come. While quantity of such datasets is considered as satisfactory enough, quality (e.g., correctness and vertical completeness) is yet to be improved [2]. Moreover, loosely coupled nature of data is posing challenge in developing applications on top of them. Therefore, there is a pressing need to leverage this data before putting them in action. To overcome the issues and fulfill the demand, we made the following contributions in this paper:

i) The definition of the opening procedure as an integral part of the change management in a public administration.
ii) The development of a methodology for generating entity types out of published datasets to model data as entities for facilitating an integrated, combined and extensible representation.
iii) The development of an algorithm and the corresponding tool for dealing with unforeseen

[1] http://innovazione.provincia.tn.it

data (along with known ones) about an entity, taking into account the semantics.

iv) Description of our experience in handling Open Big Data for building life style changing unprecedented (in the region) applications.

The paper is structured as follows: in Section 2 we describe the process we put in place with the PAT and the choices we made. In Section 3 we describe the entity type development methodology that helps creating entities. Section 3 shows the automatic creation of entities matching dataset schemas to the entity types. Section 4 provides a brief description of open big data. In Section 5, we present some applications developed on top of entities and open big data. Section 6 concludes the paper.

II. OPEN DATA TRENTINO

The Open Data Trentino[2] project was created under the push of the local government of opening their public information as expressed in the guidelines for the reuse of public data official document [3]. The process started by adapting and improving, for the local administration context, the state of the art of existing European good practises in matter of Public Sector Information (PSI)[3].

The publication process followed is a step by step, day by day, federated approach by involving the local authorities since the beginning by asking to every provincial department to open at least one dataset. Although the direct engagement of any department is time consuming it has proved of being successful and brought the desired side effect of creating enough awareness to the whole public administration and to spread the change paradigm. At the same time, we have focused at the creation of the Data as a Culture, by acting in several dissemination and educative actions internally to the institution and with a broader scope at the national and international level.

As of April 2014, the government of the PAT published about 650 datasets in a catalogue made available under the link http://dati.trentino.it/ more than a year ago, with an open license for ensuring free and unlimited use and reuse of data, representing the engagement of about 60 provincial departments. The catalogue is clustered into 13 broad categories each consists of a number of datasets, which are represented as one or more resources that are easily accessible and downloadable often in CSV and/or JSON format and occasionally in XML. Each dataset can be mapped to DCAT [4], a vocabulary designed for describing catalogues and datasets thereof for increasing interoperability. Just to cite a few of the published resources of high importance, there are provincial budget and cadastre.

The Dati Trentino portal is built over CKAN[5], an ad-hoc open source data management system started by the Open Knowledge Foundation and maintained by the CKAN community itself to which we are currently contributing. Each dataset within the catalogue has a dedicated page furnished with a good quantity of metadata facilitating its easy finding and retrieval.

At this time Open Data in the Province of Trento is well and positively recognized and more and more the formalized publication process is synonym of transparency as enabler for Open Government innovation data management.

III. OPEN ENTITY

This Section depicts an entity centric approach for modeling OGD. It describes the generation of entity types according to the data published in government data catalogues. An entity type (also called as eType) is a type or category of an entity with a set of data attributes and/or relational attributes forming the foundation of creating entities [8]. Some examples of entity types are person, location, organization and facility. An entity is a real world physical or abstract or digital thing becomes so crucial to us, worth giving it a name for referring to it. For example, Dante Alighieri (person), Trento (location), University of Trento (organization) and Trento Railway Station (facility) are entities.

In the context of this work, we have been dealing with the catalogue published by the PAT. As shown in Figure 1, we divided the entity type development in three macro-phases -- datasets survey (A), attributes

[2] http://www.innovazione.provincia.tn.it/opendata (in Italian).

[3] "Libro bianco per il riutilizzo dell'informazione del settore pubblico", EVPSI Project (2012), http://www.evpsi.org/evpsifiles/bianco_beta.pdf; Piedmont Data Catalogue, http://dati.piemonte.it, Lombardy Data Catalogue, http://dati.lombardia.it,.

[4] http://www.w3.org/TR/vocab-dcat/

[5] CKAN Platform, http://ckan.org

survey (B) and producing entity types (C) -- each of them with different macro-steps.

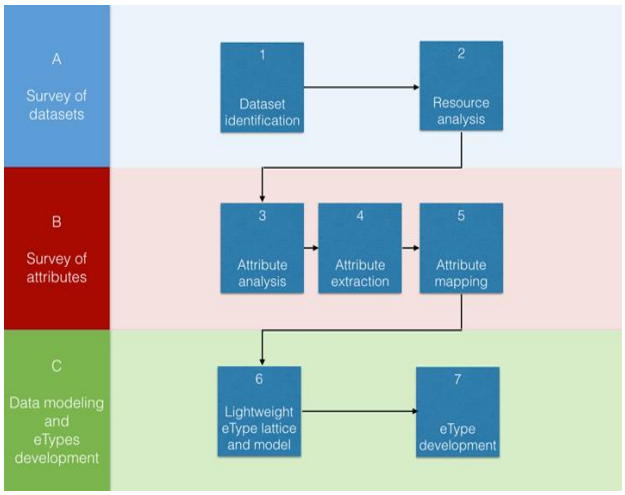

Figure 1: Modelling Open Government Data as Entities

Modeling starts with the *dataset identification* step which relies on the scenario or task at hand. For example, our scenario involves points of interest which include datasets representing among others refreshment facility (such as restaurant, pizzeria, bar, etc.), recreational facility (such as ski lift, sports ground, museum, etc.) and transportation facility (such as bus stop, railway station, cable car stop, etc.). In the *resource analysis* step a rigorous study on the structure and content of the corresponding file(s) is performed to understand the relevant resources for singling them out.

The *attribute analysis* proceeds through examining the attributes of the already selected resources. With attributes, it means column headers in the CSV files, properties of objects in the JSON files and sub-tags under repeated object tags in XML. Attribute values are also analyzed in terms of availability (e.g., always, frequently, sometimes and never present) and quality (e.g., complete data with no or occasional syntax error, partial data with or without error and relational data with or without disambiguated reference) of the data.

In the *attribute extraction* step, we differentiate between the kinds of attributes according to the data encoded in them. Some attributes are used for encoding data and some others for managing data (e.g., identifiers internal to the resource used as primary keys). The *attribute mapping* step incorporates disambiguation of the attributes proceeds through linking them to the right concepts in the knowledge base.

The *data modelling* step leads to understanding an entity type from the attributes extracted in the previous step and finding it (if exists) or a suitable parent (if created newly) in the already existing *entity type lattice*, a lightweight ontology (see [5]) formed with the concepts of the entity types. In the *entity type development* step, we produce a specification of an entity type defining all possible attributes, their data types (e.g., string, float) and meta attributes such as permanence (e.g., temporary, permanent), presence (e.g., mandatory, optional) and category (e.g., temporal, physical).

While producing entities of a given entity type, mandatory and optional attributes are filled in with data, which are semantified and disambiguated wherever applicable. In fact data for an entity can come from multiple resources. Through semantification, we facilitate the integration of loosely coupled data. In the case of unavailability of a mandatory attribute in the possible resources, we signal it to the data provider as *pro-sumers* (see [6]) and do not allow the creation of the corresponding entity unless all necessary data are present. This is how we can improve the vertical completeness.

IV. Open Data Rise

Open Data Rise[7] (ODR) is a web application to refine and semantify OGD. The framework employs the entity-centric model and extends Open Refine[8], a famous tool for cleansing messy data. The semantification pipeline provided by ODR consists of seven steps shown in Figure 2.

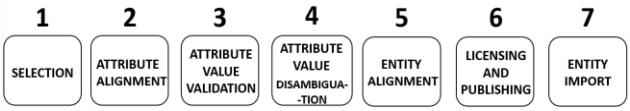

Figure 2: ODR Semantification Pipeline

During selection step the framework allows the user to select a dataset from any CKAN compliant repository. For catalogs previously analyzed by Ckanalyze[9], a component we developed, repository statistics are also shown.

On the *attribute alignment* step the framework matches columns in the source data set to a predefined set of entity types. A simplified example is presented in Figure 3.

[7] https://github.com/opendatatrentino/OpenDataRise
[8] http://openrefine.org/
[9] https://github.com/opendatatrentino/CKANalyze

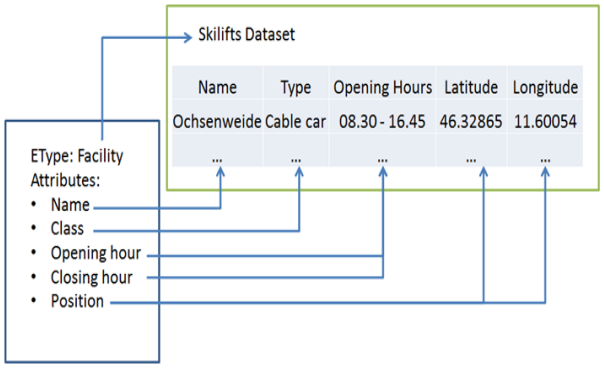

Figure 3: An example of a simplified schema matching

This step includes finding the appropriate entity type for a given data set and mapping its attributes with column headers in the dataset. Once dataset schema has been determined, during *attribute value validation* step the user can adapt the dataset to the schema, exploiting *OpenRefine* data cleansing capabilities.

Successive *attribute value disambiguation* step employs Natural Language Processing techniques for enriching dataset content by linking names to known entities (such as *Dante Alighieri*, *Florence)* and words to concepts (such as *male*, *city).*

Within *entity alignment* step the framework considers rows in the dataset as entities, i.e. real instances. The goal of this step is to either update existing matching entities in entity storage or to create new entities with values from the source dataset.

The license of entities to publish can be determined during *licensing and publishing* step. Finally, during *entity import* step updates to entity storage are committed and enriched dataset is published to CKAN.

V. Open Big Data

As part of OGD initiative of the PAT, we also focus on the problem of data explosion and the consequent need of having fast and scalable solutions for storage and analysis. Even though currently the size of the complete published Open Data datasets is not relevant in terms of storage requirement, we estimate the trend for growth will be up to hundred times per year, easily reaching TB of data by 2018. For instance the Trentino portal already have sensors based datasets, such as weather, traffic sensors and real time position of the city buses that already provide few GB of data per day. Due to this very nature, they pose challenge in using traditional relational database management systems to handle them and at the same time appear as a problem to be dealt with the big data technologies such as Apache Hadoop[10] and several others. A kind of big data generated by various actors including government agencies and companies and published as open data is called *open big data*. The interest here is focused on producing a BigData platform directly integrated with the Open Data portal and able to provide useful analytics, historical analysis in reasonable time.

VI. Applications

Open entity as well as open big data exploitation resulted to the development of applications which often appear as innovations to the citizens. This is because they offer services that are either novel or come up with better results in comparison with the contemporary ones. Moreover, open data propel innovations help devising novel applications and services [7].

In line with this consideration, as shown in Figure 4 and Figure 5, we developed an application running on top of open entities helps finding points of interest including restaurant, pizzeria and bar with opening hours and bus stop, cable car stop and railway station with timetable and ski lift, ski rental and ski school with timetable.

Figure 4: Faceted navigation for finding points of interest

Figure 4 sketches faceted navigation. For example, when user is in a ski lift location, selecting a point of interest category or a subcategory shows the corresponding results on the map. Figure 5 shows semantic navigation. We call it semantic navigation as it exploits semantic relations like part-of while exploration based search proceeds. To provide an

[10] http://hadoop.apache.org/

example, when user is in a museum, it can show, within a radius of, e.g., 500 m or 1 km, all kinds of points of interest which are also located in the place where the museum itself is.

Figure 5: Semantic navigation for finding points of interest

OpenBus [11] shown in Figure 6 is an application developed using open big data. This application approximates real time (geographical) positions of the city buses in all the routes in Trento, based on 2 minutes temporal snapshots of their registered geo-location.

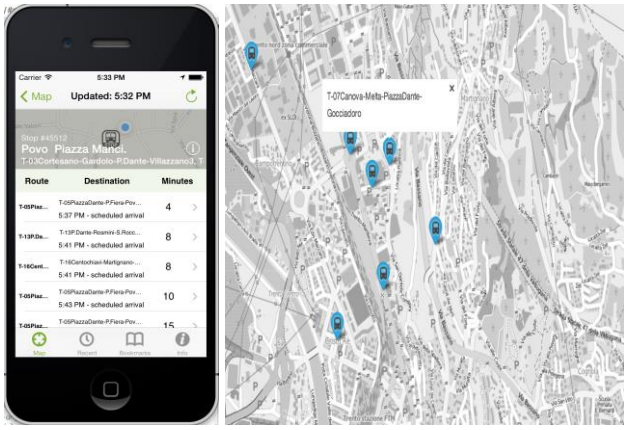

Figure 6: Open bus

It employs big data analytics tools and techniques in analyzing data generated by on-board sensors to predict delays on the basis of the current position of the buses.

VII. CONCLUSION

In this paper, we proposed an approach for generating entity leveraging open government data. The result that we obtained in Trentino is promising and helpful in the diffusion of the *data culture* to the public administrations. We are building an entity based infrastructure that by design facilitates consistency in data representation and management and as such enables the re-use of public sector information [4]. This entity centric data representation and the infrastructure as a whole can be considered as an input to the W3C Data on the Web Best Practices Working Group to provide guidance to data publishers. Our future work involves the creation of a broader data ecosystem. Communities, universities and other actors will be able to link their data using Open Data Rise. Moreover, we will implement crowdsourcing techniques for improving the quality of the existing data.

[11] http://livemap-bigdata.trentorise.eu/

ACKNOWLEDGMENT

Financial support for this project is provided by the Autonomous Province of Trento. The authors wish to thank Dr. R. Cibin, Dr. L. Paolazzi and Dr. L. Ruzzene for their valuable contribution at the realization of the project.

REFERENCES

1. Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing.
2. C. Bohm, M.Freitag, A.Heise, C.Lehmann, A.Mascher, F.Naumann, V.Ercegovac, M. A. Hernandez, P.Haase, M.Schmidt. Gov-WILD: integrating open government data for transparency. In Proceedings of WWW, pages 321–324, 2012.
3. Delibera Giunta Provinciale 2858/2012. http://www.innovazione.provincia.tn.it/binary/pat_innovazione/notizie/Lineeguida_21dicembre_def.1356705195.pdf
4. European Parliament and Council (2003). Directive 2003/98/EC of 17 November 2003 on the re-use of public sector information. OJ L, 345, 90.
5. Giunchiglia, F., Marchese, M., Zaihrayeu, I.: Encoding classifications into lightweight ontologies. Journal of Data Semantics (2007)
6. Charalabidis, Y., Loukis, E., and Alexopoulos, C.: Evaluating Second Generation Open Government Data Infrastructures Using Value Models. Hawaii IEEE International Conference on System Sciences, 2014.
7. Chan, C.M.L. 2013, From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data. Hawaii IEEE International Conference on System Sciences, 2013.
8. Farazi, F.: Faceted Lightweight Ontologies: A Formalization and Some Experiments. PhD Thesis, International Doctoral School in ICT, University of Trento, Italy, 2010.