Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

TRENTINO GOVERNMENT LINKED OPEN GEODATA: A CASE STUDY

Pavel Shvaiko, Feroz Farazi, Vincenzo Maltese, Alexander Ivanyukovich, Veronica Rizzi, Daniela Ferrari and Giuliana Ucelli

September 2012

Technical Report # DISI-12-032

A revised version of this technical report has been published in the proceedings of the ISWC Conference 2012.

Trentino government linked open geo-data: a case study

Progetto: Semantic Geo-Catalogue 2

Reference persons:

Informatica Trentina: Pavel Shvaiko

DISI Trento: Feroz Farazi, Vincenzo Maltese, Veronica Rizzi

Trient Consulting: Alexander Ivanyukovich

Segreteria SIAT, PAT: Daniela Ferrari, Giuliana Ucelli

Our work is settled in the context of the public administration domain, where data can come from different entities, can be produced, stored and delivered in different formats and can have different levels of quality. Hence, such heterogeneity has to be addressed, while performing various data integration tasks. We report our experimental work on publishing some government linked open geo-metadata and geo-data of the Italian Trentino region. Specifically, we illustrate how 161 core geographic datasets were released by leveraging on the geo-catalogue application within the existing geo-portal. We discuss the lessons we learned from deploying and using the application as well as from the released datasets.

1 INTRODUCTION

Our work is settled in the context of the public administration (PA) domain. It gathers applications with a variety of constraints, interests and actors including citizens, academia and companies. Within PA, data can come from different bodies, can be produced and stored in different formats and can have different levels of quality. Thus, such heterogeneity has to be addressed, while performing various data integration tasks.

We describe how, within the semantic geo-catalogue application [7, 18], the Autonomous Province of Trento (PAT) has published some of its core geo-data accompanied with the corresponding metadata following the open government data (OGD) and the linked open data (LOD) paradigms. The goal is to experiment in practice with the realization of such paradigms to obtain insights on how the services offered by the PA can be improved and the above mentioned heterogeneity can be tackled more efficiently.

The need for coherent and contextual use of geographic information between different stakeholders, such as departments in public administrations, formed the basis for a number of initiatives aiming at sharing spatial information, e.g., the INSPIRE (www.ec-gis.org/inspire/). See, for instance the work in [19, 22]. Even though the publication of LOD is not required by the INSPIRE directive [1] our approach can be considered as a novel good practice to this end. In fact, in parallel with the standardization and regulation effort, the implementation of INSPIRE should take into account the linked data principles, since they facilitate data harmonization. For instance, the issue is to identify the most relevant vocabularies for RDF representation of the INSPIRE metadata elements. Also geo-data, modeled as INSPIRE themes, can be represented as RDF triples in order to facilitate its discovery and future re-use. Within the European Commission, the process has already started, for example for the INSPIRE data theme "addresses" specification which was used as a basis to model the "Address" class of the Core Location Vocabulary of the Interoperability Solutions for European Public Administration (ISA) program (tinyurl.com/72538jm).

In turn, the OGD paradigm encourages governments to publish their data in an open manner (from both technical and legal perspectives) to foster transparency and economic growth (through data re-use). The theme of linking open government data gains more interest as it aims at simplifying data integration [27], e.g., by providing explicit links in advance to other relevant datasets. Consider for example the US (www.data.gov) [5] and UK (data.gov.uk/) [16] initiatives.

Our work includes:

- Description and analysis of concrete problems in the eGovernment domain;
- Details of the implementation and usage scenarios of a semantic application that manages the released 161 core geographic datasets;
- Lessons learned from deploying and using the application and the datasets.

The argumentation is as follows. Section 2 provides the problem statement. Section 3 articulates the approach adopted. Sections 4-6 present the solution realized. Section 7 outlines the related work. Section 8 discusses the lessons learned. Finally, Section 9 reports on the major findings.

2 THE APPLICATION SETTING

Our application domain is eGovernment, i.e. an area of application for ICT to modernize public administration by optimizing the work of various public institutions and by providing citizens and businesses with better (e.g., more efficient) and new (that did not exist before) services. More specifically, we focus on geographic applications for eGovernment. At the European level, the INSPIRE directive aims at creating the framework for sharing spatial information by providing the respective rules leading to the establishment of such a framework. At the national level, DigitPA has produced the so-called Repertorio Nazionale Dati Territoriali (RNDT, www.digitpa.gov.it) that constrains further the INSPIRE requirements for Italy. At the regional level these developments have been subsequently put in practice by requiring the existing systems to evolve in the respective directions.

2.1 The context

One of the key components of the INSPIRE architecture is a discovery service, that ought to be implemented by means of the Catalogue Service for the Web (CSW, www.opengeospatial.org/standards/cat) - a recommendation of the Open Geospatial Consortium (OGC) - which is often realized within a geo-catalogue. See Fig. 1 for an overview.

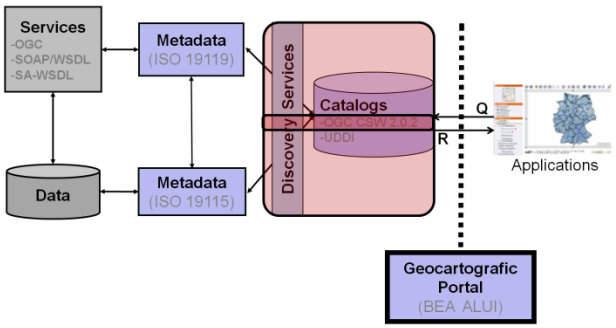

Fig 1. – Discovery services

Specifically, geo-data (e.g., in shape files) is described by metadata conforming to the ISO19115 standard. In turn, it can also be made available through services, such as OGC WMS (web map service) for map visualization or WFS (web feature service) for downloading maps (features), which are described by metadata conforming to the ISO19119 standard. Metadata is handled through a catalogue service, such as OGC CSW. The catalogue can be accessed either through applications or a web portal. We focus only on the latter.

Essentially, the geo-catalogue offers a standard mechanism to classify, describe and search information on geo-data and geo-services conforming to the above mentioned standards. There are several implementations of the CSW-based geo-catalogue, e.g., Deegree (www.deegree.org/) and GeoNetwork (geonetwork-opensource.org/). We have used GeoNetwork Open source (version 2.6). Its major functionalities include:

- **Metadata management**: search, add, import and modify metadata;
- **User and group management**: import users, their role, transfer metadata ownership;
- **System configuration**: use various languages and harvest metadata from remote sites.

2.2 Towards Trentino ODG

The benefits of opening government data have been recognized at the regional level, namely in terms of:

- increased transparency for the PA;
- potential economic growth through data reuse, and hence, creation of new business opportunities;
- potential increased participation of citizens in PA.

Nevertheless, a critical mass has not been created yet to launch a transversal initiative in the data.gov.uk spirit. Thus, we have followed a *low hanging fruits first* approach by postponing a global strategy formulation and a road mapping activity to a later stage, though by taking already into account the available studies in these respects [15, 24].

Operationally, we have introduced the task of experimenting with open government data within an ongoing project, which is on realizing a semantic geo-catalogue [7, 18]. This choice was made to rapidly create practical evidence on the expected benefits with reduced costs. Thus, we have done a vertical experimentation by adapting the available geo-catalogue system, rather than by creating a new dedicated one.

3 THE APPROACH

The OGD paradigm fosters openness in both legal and technical directions. With respect to the legal openness, data should be published under a suitable license, such that third parties could freely use, reuse and redistribute it. The Open Knowledge Foundation (OKF, opendefinition.org) community provides a summary for such licenses. To this end, under the recent regional deliberation n. 195/2012, the PAT formally decided to adopt Creative Common Zero (public domain) license to release 161 of its geographical core datasets. They include: bicycle tracks, administrative boundaries, ski areas and CORINE land cover.

With respect to the technical side, Trentino has been the first administration in Italy at the regional level that published its data following the linked open data principles, also known as a 5-star rating system [2]. Specifically, we followed a standard publishing pipeline (similar to the one proposed in [12]) constituted by the following sequential phases:

Step 1: Conversion of raw data in RDF. Data and metadata of the identified datasets were automatically converted in RDF. Data was available in shape files (SHP) and metadata in XML. Data was pre-processed with the GeoTools (www.osgeo.org/geotools) to produce XML. Both data and metadata were then processed with a SAX Parser (www.saxproject.org/) to extract information that were finally given in input to the Jena tool (jena.apache.org/) to produce the corresponding RDF.

Step 2: Linking. To favour interpretation of the terms used and interoperability among different datasets, data and metadata are linked to external vocabularies. The high quality of links was guaranteed by validating them manually. This has been done at the level of classes, entities and their attributes. Even if this is clearly somewhat time consuming in general, in our case this is motivated by the limited number of datasets and because of the unsatisfactory quality of the links that we obtained by using the existing linking facilities, such as Google Refine [12] and Silk [25].

Step 3: Sharing. The RDF data produced is made available for sharing. Our datasets are published on a web server and can be downloaded from the Trentino geo-portal. For each class (e.g., river, bicycle track) a different RDF file can be accessed.

Step 4: Evaluation. RDF data is evaluated by means of a developed mash-up. This has been done through the use of DERI pipes [13] that allowed fast prototyping of mash-ups using different data sources. We have also run a workshop with the participation of the public administration, academia and industry to share and discuss the experience gained with the exercise (www.taslab.eu/trentino-open-data-primi-risultati).

4 CONVERSION AND LINKING

Within this task, both metadata and data of the 161 selected geographic datasets were automatically converted into RDF and manually linked to relevant vocabularies. To facilitate discovery and re-use, each dataset - corresponding to a different geographical feature - was converted into a different RDF file.

Metadata was initially available in the XML format. For the conversion of XML metadata into RDF, existing tools usually rely on a rule file providing the mapping between the source XML and the target RDF objects [26]. But, the work following this line is often limited by the non-trivial requirement of learning a tool specific rule language and the unsatisfactory quality of the generated RDF. As an alternative we used a SAX parser to retrieve metadata from XML files. Among the widely used tools for parsing XML, we chose SAX over DOM (www.w3schools.com/dom/dom_parser.asp) because of the high memory consumption limitation of the latter.

Geo-data was given in shape files. GeoTools, an open source java library, was used to convert them into XML, which were then parsed using SAX to retrieve data.

Both metadata and data were then fed to Jena to produce RDF.

4.1 Geo-metadata conversion

With the emergence of the Linked Open Vocabulary LOV, labs.mondeca.com/dataset/lov/) several vocabularies are being published and similar ones are being grouped together. As a result, finding a suitable vocabulary for publishing a specific dataset in RDF has become easier. In case of unavailability of a suitable one, users can eventually propose a new vocabulary. However, in order to maximize interoperability among datasets it is important to select a vocabulary among those with wider consensus. For this reason, we have encoded geographic metadata - originally provided following the ISO19115 standard - using Dublin Core

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dcmibox="http://dublincore.org/documents/dcmi-box/"
    xmlns:dc="http://purl.org/dc/elements/1.1/" >
<rdf:Description rdf:about="http://www.territorio.provincia.tn.it/geodati/p_tn:piste_ciclabili">
    <dc:language>it</dc:language>
    <dcmibox:westlimit>10.41</dcmibox:westlimit>
    <dcmibox:eastlimit>11.97</dcmibox:eastlimit>
    <dcmibox:southlimit>45.60</dcmibox:southlimit>
    <dcmibox:northlimit>46.60</dcmibox:northlimit>
    <dc:identifier>http://www.naturambiente.provincia.tn.it/</dc:identifier>
    <dc:format>shp</dc:format>
    <dc:rights>Uso limitazione: nessuna limitazione. Altri vincoli: Dato pubblico</dc:rights>
    <dc:title>Piste ciclabili</dc:title>
    <dc:creator>Dipartimento Risorse Forestali e Montane</dc:creator>
    <dc:version>1.0</dc:version>
    <dc:date>2008-09-26</dc:date>
  </rdf:Description>
</rdf:RDF>
```

Fig. 2 – Fragment of encoding geo-metadata in RDF

(DC, dublincore.org/documents/dces/) and DCMI-BOX (dublincore.org/documents/dcmi-box/) standard vocabularies. See example in Fig. 2.

In particular, we have focused on those metadata elements which fall in the intersection of INSPIRE/ISO Core metadata and DC. They were grouped under a resource, which was given a URI generated by appending the file identifier, e.g. *p_tn:piste_ciclabili* metadata attribute to the namespace URI for the Trentino datasets (www.territorio.provincia.tn.it/geodati/)

The metadata resource language, online locator, distribution format, use limitation, title, responsible organization, version and creation date were (obviously) mapped to *dc:language*, *dc:identifier*, *dc:format*, *dc:rights*, *dc:title*, *dc:creator*, *dc:version* and *dc:date*, respectively; the geographic bounding box attributes west bound longitude, east bound longitude, south bound latitude and north bound latitude were mapped to *dcmibox:westlimit*,*dcmibox:eastlimit*, *dcmibox:southlimit* and *dcmibox:northlimit*.

4.2 Geo-data conversion

An example of how geographic data from shape files was selectively published in RDF can be found in Fig. 3. To express the geographic position of the features, the UTM coordinate system was preserved. New terms were created only in case not suitable candidates were available in the standard vocabularies [10]. Specifically, we have created the length, area, perimeter and polyline terms. When available, we have specified the length of the features modeled as polylines and the area and perimeter of the features modelled as polygons.

Geometric objects that are found in data are points, polylines and polygons. A point consists of latitude and longitude geographical coordinates. A polyline shape is formed by a set of points, with two consecutive points that are connected by a line. A polygon shape is formed by a set of points, with two consecutive points that are connected by a line and with the first point and the last point that are the same. We have encoded all the points of the polylines and polygons in RDF.

4.3 Linking

With this step we have linked our RDF to some of the most highly connected hub datasets from the linked open data cloud. As it can be seen from Fig. 3, this has been done through the *owl:sameAs* OWL association. To ensure a high accuracy, the links between the resources were established manually and it took one working day.

In line with the *low hanging fruits first* approach that we have followed, we have started with DBPedia (dbpedia.org) and Freebase (www.freebase.com/). In fact, being among those with higher connection with other datasets, they guarantee a high level of reusability and interoperability. Despite they are not domain specific, they also have a broad coverage in our domain of interest.

As next step we will link the RDF data to geographic datasets, such as GeoNames (www.geonames.org). Also dataset ranking mechanisms, such as in [20], can be employed. As a matter of fact, we did not include GeoNames from the beginning as it lacks of features that were central to the evaluation (Section 6), such as bicycle tracks that at the moment is also one of our most downloaded datasets.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:geontology="http://www.territorio.provincia.tn.it/geodati/ontology/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#" >

  <rdf:Description rdf:about="http://www.territorio.provincia.tn.it/geodati/resource/piste_ciclabili">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <owl:sameAs rdf:resource="http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000000428308"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://www.territorio.provincia.tn.it/geodati/resource/piste_ciclabili/529">
    <geontology:length rdf:datatype="http://www.w3.org/2001/XMLSchema#double">1445.8484810675</geontology:length>
    <rdfs:label xml:lang="it">Mori - torbole</rdfs:label>
    <rdf:type rdf:resource="http://www.territorio.provincia.tn.it/geodati/resource/piste_ciclabili"/>
    <rdfs:label xml:lang="it">529</rdfs:label>
    <geo:geometry rdf:resource="http://www.territorio.provincia.tn.it/geodati/resource/piste_ciclabili_529"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://www.territorio.provincia.tn.it/geodati/resource/piste_ciclabili_529">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <geontology:polyline>646339.346896746,5082179.74045936
                         646329.929020191,5082161.84683082
                         ...
                         645576.090351533,5081173.94569307
                         645575.851739799,5081173.68539361
    </geontology:polyline>
  </rdf:Description>
</rdf:RDF>
```

Fig. 3 – Fragment of encoding geo-data in RDF

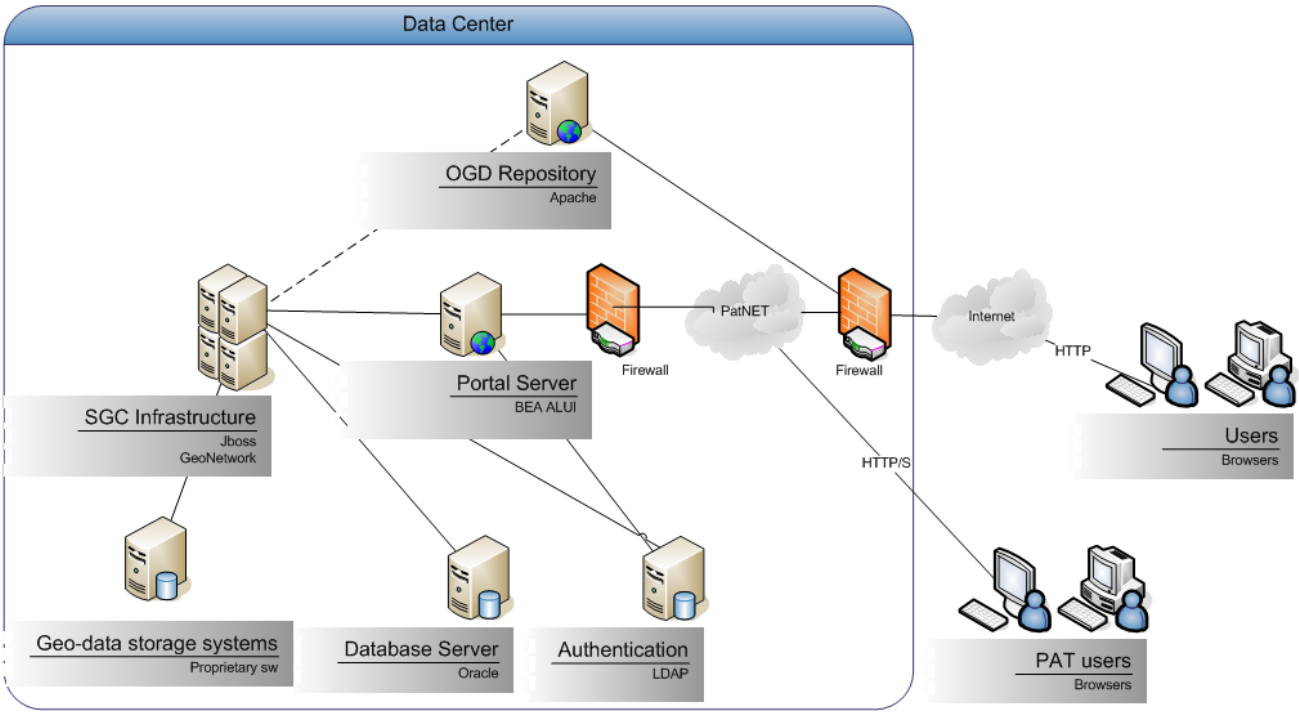

Fig. 4 – System architecture

5 SHARING

The INSPIRE directive indicated quality of service criteria to be respected and monitored by the implementing systems:

- **Performance:** to send one metadata record within 3 seconds;
- **Availability:** service available by 99% of time and no more than 15 minutes downtime per day during working hours;
- **Capacity:** 30 simultaneous service requests within 1s.

Other requirements we had to comply with include:

- coherent view among other geo-related services offered by the PAT,
- centralized user authorization and authentication using standardized mechanisms,
- usage of standard architectures and interfaces for inter-system communications.

To satisfy these requirements, the system architecture shown in Fig. 4 was implemented. It involves the following main software components:

OGD repository is a web-based component responsible for the access to the datasets released. It is based on the Apache web-server.

Portal server is a basis of the geo-portal of the PAT and is an umbrella for all projects of the province dealing with geographical information. It groups them together and serves as a single entry point for citizens and companies. Portal server is based on the BEA ALUI proprietary software solution.

Geo-catalogue (SGC) infrastructure is responsible for the access and management of geo-information (metadata and data). It is based on GeoNetwork open-source software personalized for integration with the existing proprietary software of the PAT.

Geo-data storage systems are back-end systems that store geo-data in various formats (e.g., shape files). These systems are internal systems of the PAT.

With reference to Fig. 5 in the following we describe how information can be accessed by using the Trentino geo-portal (www.territorio.provincia.tn.it). First of all, in order to access the geocatalogue (*Ricerca nel Geo-catalogo*), the user must select SIAT (*Sistema Informativo Ambiente Territorio*) from the main menu. Users can issue queries by typing them in the search box (1) and by clicking on the corresponding search button (2). Queries can be simple, such as bicycle tracks, or more complex ones, such as Trentino mountain hovels reachable with main roads. These are semantically expanded (see [7] for a description of how this is done) and executed against the existing metadata records. Search results are shown as a list of datasets below the search box. The header on top of the list shows the total number of the datasets found and the number of datasets displayed on the current page. Each dataset is presented on the results page with its title, contact information (e.g., “department of forest resources and mountains”), keywords and description. Possible operations that can

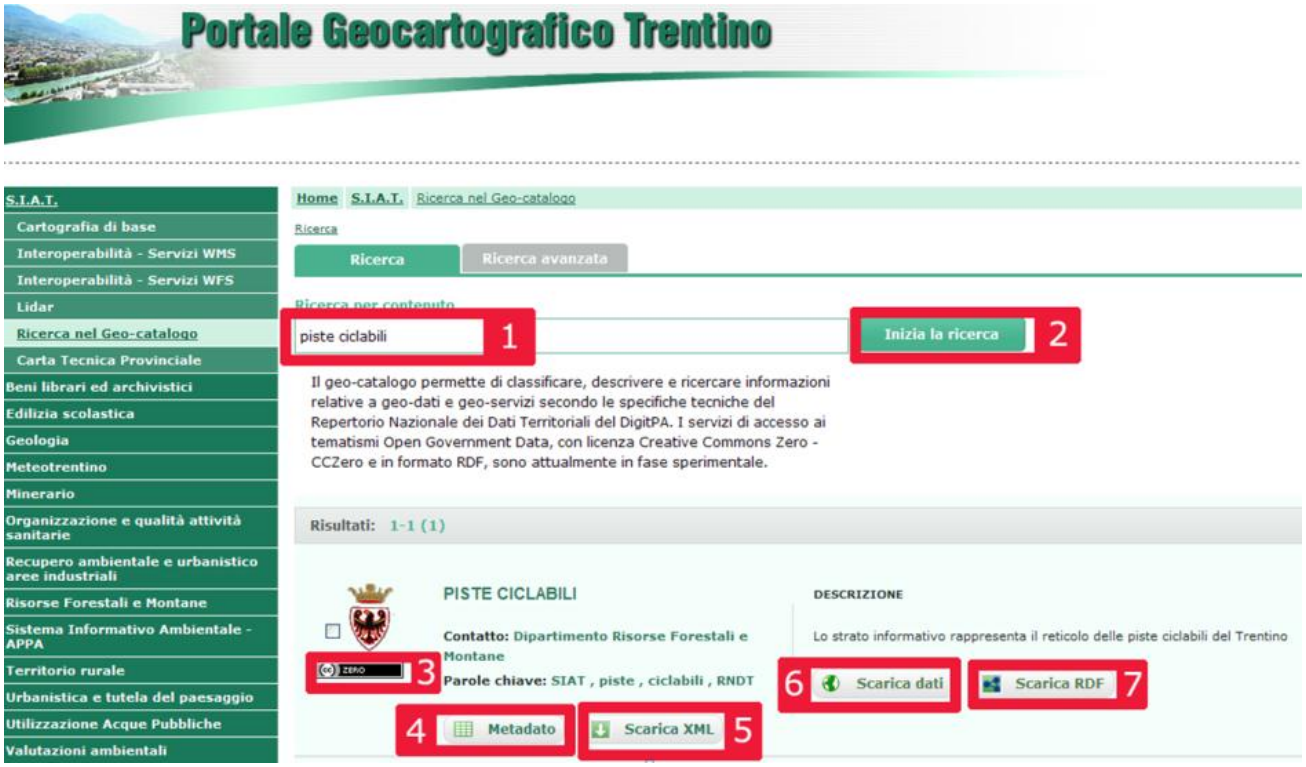

Fig. 5 – Search results

be performed on the dataset include: (4) display the geo-metadata; (5) download the geo-metadata in XML format; (6) download the raw geo-data (in a ZIP package); (7) download the dataset in RDF (see Section 4). The icon (3) indicates that the dataset is released under the Creative Commons Zero license (CC0).

6 EVALUATION

To evaluate our datasets we have built a mash-up application (http://sgc.disi.unitn.it:8080/sgcmashup/). It enabled us to observe the usefulness of the published geo-data in linking and accessing different datasets. The purpose of this application is to support the following scenario:

> *Robert is in a summer trip to Trento cycling along the bicycle path between Trento and Riva del Garda. Once he arrived in the lakefront region of the Mori-Torbole bicycle track, he is fascinated by the splendid natural beauty of the lake and the panoramic beauty of the mountains, which made him interested to know more about the panoramic views of the other parts of the bicycle track and the nearby hotels to stay there for some days. Cycling in the summer noon made him thirsty. Hence he is eager to know the location of the drinking water fountains in the vicinity of the bicycle track.*

Fig. 6 provides a snapshot of the mash-up application supporting this scenario. Streams (e.g., Adige), bicycle tracks (e.g., *Mori-Torbole 507*) and bicycle track fountains are shown on the left as a list of check boxes, where the numbers to the right of the tracks represent the identifiers of the track parts which constitute the whole track. Selected streams, bicycle tracks and fountains are displayed using Google Maps as polygons, polylines and markers, respectively. By clicking on a bicycle track it is possible to visualize a set of images of the nearby hotels and panoramic views. We have collected images from Flickr and we have gathered information about fountains from Open Street Map through LinkedGeoData (linkedgeodata.org/).

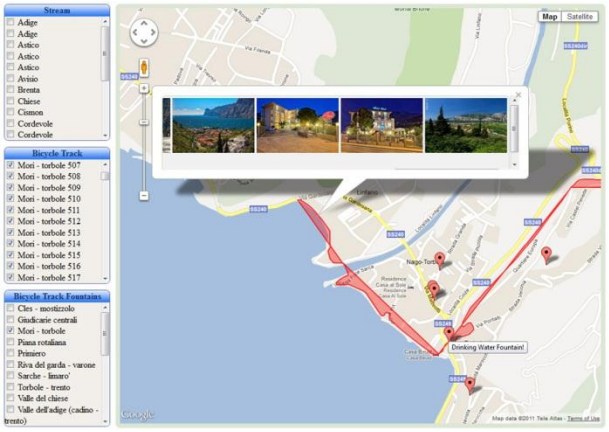

Fig. 6 – The developed mash-up application

To combine information from different RDF resources, we have used the DERI pipes tool [13]. The development of this mash-up on top of the linked geo-

data took a short time (about 4 working days) compared to the time required if we were to develop the same mash-up without using semantic technologies. It has required less time because, among others:

- it has avoided the need for solving data heterogeneity issue as linked data are published in RDF or RDF compatible format;
- it has overcome the spatial restriction on data, e.g., necessity to have all data in the same database, as it has worked simply by referring to the dataset URLs;
- including a new dataset to an application is less time consuming because of the open (known) data format and ease of access to data through URLs.

Finally, we have asked a local start-up company, SpazioDati.eu, to use the released datasets and in one week the company was able to design a business idea suitable to be presented at the regional workshop (www.taslab.eu/trentino-open-data-primi-risultati) dedicated to the release of the datasets. As a result, at the workshop they presented the Tindes, a naturalistic index computed for the Trentino restaurants together with a mobile app and widget implementations. Overall, 32 PAT datasets were reused and mixed with 9 Open Street Map datasets. This has provided additional evidence of the usefulness of the released datasets and the possibility to build new business opportunities using them.

7 RELATED WORK

In creating and publishing government data, the contribution of both the public administrations and universities is noticeable. In this section, we review the related work and compare it with the approach we followed along two lines: (i) open government data and (ii) publishing open data.

7.1 Open government data

Governments are becoming more and more active w.r.t. OGD. Specifically concerning geospatial data, the UK government has decided to publish them following the INSPIRE Directive using open standards, e.g., RDF for representation, SPARQL Endpoint for exposing, DCMI (Dublin Core Metadata Initiative) vocabulary for annotation and GML (Geography Markup Language, www.opengeospatial.org/standards/gml) for representing geographic features. Basically, the use of a SPARQL Endpoint for exposing data allows the Semantic Web search engines, e.g., Sindice (www.sindice.com), Swoogle (swoogle.umbc.edu) and Watson (watson.kmi.open.ac.uk), to discover, crawl and index the RDF data which in turn helps increasing the visibility of the data itself. Ordnance Survey (www.ordnancesurvey.co.uk), the national mapping agency in the UK, spearheaded the publishing of geospatial information as part of the linked data [9].

In Portugal, the Geo-Net-PT [11] dataset was created at the University of Lisbon to support applications requiring national geographic information. This dataset is in RDF and it is linked to Yahoo!GeoPlanet (developer.yahoo.com/geo/geoplanet). Standard vocabularies were used including DCMI for metadata and WGS84 vocabulary for geographical coordinates. This dataset is also used as geospatial ontology. A SPARQL Endpoint is provided for querying it. The quality of this work is significant.

In Spain, the GeoLinked Data [3] initiative at the University Politecnica de Madrid has contributed to bringing Spanish geographic and statistical information to the linked data cloud. They have dealt with the data sources owned by the Spanish National Geographic Institute (IGN-E, www.ign.es) and Spanish National Statistical Institute (INE, www.ine.es). Their dataset is linked to GeoNames and DBPedia. For the representation of the statistical (e.g., unemployment rate), geometrical (e.g., shape) and geo-positioning (e.g., geographical coordinates) information, Statistical Core Vocabulary (SCOVO, vocab.deri.ie/scovo), GML and WGS84 vocabularies were used, respectively. To the best of our knowledge, similarly to Geo-Net-PT, it did not go to production.

In Italy, many communities promote OGD activities. For instance, DataGove.it aims at promoting an open and transparent government in Italy. Trentino Open Data (www.trentinoopendata.eu) aims to sensitize public awareness of open data issues starting from the Trentino region. Moreover, in Italy many public administrations, for instance, the Piedmont region (dati.piemonte.it), are working to publish their datasets following the principles stated by OKF. However, at the time of writing, to the best of our knowledge the coverage of their published RDF datasets is quite limited (only 3 features: schools, municipalities and provinces) and no links are provided to any external datasets.

7.2 Publishing open data.

In the following we compare the way in which we have published the open data versus alternative approaches from the state of the art.

Conversion. In [12] data conversion was accomplished with the condition that the dataset had to be published in the Dcat (vocab.deri.ie/dcat) format. This is a strong limitation since in case data is not already in this format there are no tools to automatically convert other formats (e.g., CSV, XML) into Dcat. As a result, here data conversion was not automated.

Linking. In our work the high quality of links was guaranteed by validating them manually. In GovWILD [4] links were established automatically with specifically developed similarity measures. In Midas

[14], data about government agencies were matched by using government data extracted from documents. In [12] the alignment was done semi-automatically with Google Refine. Despite some studies show that their accuracy is good, one drawback of this and similar tools stands in the necessity to learn a specific language to handle expressions. These languages are used to specify the information which is necessary to discover the links between source and target datasets. This information includes URLs and candidate entity classes (e.g., river) and it is stored into a link specification file. Another limitation stands in the fact that they only act syntactic matching between the names of the classes. Therefore, they are unable to discover equivalent classes whose names are synonyms (e.g., stream and watercourse) or classes which are more specific (e.g., river is more specific than stream), though some ontology matching techniques can be of help here [6, 17, 23].

Sharing. We have published our datasets by making them available on a web server. What we have done is similar to what has been done previously with GeoWordNet [8]. Alternative approaches include the usage of a SPARQL Endpoint (see, e.g., in [3, 21]). In particular, in [21] along with the experiments on GeoSPARQL and geospatial semantics with the U.S. Geological Survey datasets, they show the corresponding images of the SPARQL output. In [5, 12] data sharing is enabled by loading files into CKAN (http://ckan.org/).

Evaluation. We have evaluated the generated RDF linked data with DERI pipes [13] by building a mash-up application. DERI pipes have the advantage of being open source as opposed to the proprietary software alternatives like SPARQLMotion (www.topquadrant.com/products/SPARQLMotion.html).

We did not have to handle enormous quantities of data. For data intensive applications, Hadoop is often used. For instance, in [4, 14], JSON, Jaql query language and Hadoop are used to provide citizens with information about U.S. government spending.

8 LESSONS LEARNED

Here we summarize the lessons learned from deploying and using the application as well as from the release of the datasets. These lessons are articulated along the four steps (Section 3) of the approach that we have followed:

Conversion. There is still an open question with URIs, namely which patterns to adopt. The geo-catalogue system uses by default universally unique identifiers for its records. For example, bicycle tracks correspond to *7B02F1D1-01C3-1703-E044-400163573B38*, while PA would want they were self-explanatory. Thus, an approach to URI design is still to be devised and implemented. The experimentation was useful anyhow to this end, since it has increased awareness in PA that this is not a minor detail, and that URIs enable people and machines to look them up and to navigate through them to similar entities. This is especially important for the core geographic information, which is meant to last in time, and thus, should represent precise and stable reference in order to facilitate its future reuse.

Linking. This is an important process, since it results in connecting the released datasets to the linked open data cloud, and hence, additional information can be discovered and integrated more easily. Experience with existing linking research tools revealed that they are still not yet flexible and precise enough, hence, manual process was preferred.

Sharing. We already had a basic version of a catalogue for geo-data with some metadata conforming to the respective standards (Section 2.1). We asked public administration to improve the quality of metadata and this was completed in a reasonable amount of time. This clearly facilitated the process of publishing the selected datasets. Releasing the datasets under the Creative Commons Zero license was well received by various communities with various re-launches of the news (epsiplatform.eu/content/trentino-launches-geo-data-portal). The Trentino geo-portal, being a single point of access to the geographic data, was also perceived as an appropriate place to publish the datasets. However, if this approach worked well in the context of the first experimentation, it does not scale and would create confusion, when other areas, such as statistics, culture, or tourism will start releasing their datasets.

Evaluation: The internal mash-up development and a workshop with PA, academia and industry (Section 6) has indicated that the approach adopted was a useful tactic. Local companies have perceived the value of data released by PA and would be interested in having a service for the programmatic access to the data with clear service level agreements (e.g., to have up-to-date data). This would allow them to rely on such a service and build their own applications on top of it. Also the possibility of having a feedback loop with citizens or companies in a web 2.0 fashion, signalling that some data is not precise or complete enough have to be respectively treated.

Within this experimentation we have released about 40% of the core geographic datasets of PAT. We have noticed that individuating, understanding them as well as providing metadata for them is an effort requiring collaboration of the departments owning and maintaining the respective data. We think that such datasets are of high importance, since geographic information provides a basic layer for many location-based services. The most downloaded datasets so far are administrative boundaries, bicycle tracks, and monitored rivers. With this *low hanging fruits first* approach we have managed to gain a momentum, such that an overall strategy for releasing linked open government data of Trentino should be devised briefly.

This exercise has also revealed some expectations towards the evolution of the linked open data field. For example, it has emerged the need for technology selection for the production environment to handle RDF. Comparative and convincing surveys with evaluation details are still missing that would allow for informed decision making. There is a need for instruments that support the linked data lifecycle, for example, for monitoring (and improving) the quality of data and on performing in a more automated fashion data linking and reconciliation with quality levels known in advance.

9 CONCLUSIONS

We have presented our experimental work on releasing some of the Trentino government geo-data and geo-metadata following the open government data and linked open data paradigms. Creative Commons Zero license was adopted for the release of the datasets identified. RDF has been used for representing fragments of both geo-data and the respective metadata. We have used well-known standards and specifications including Dublin Core for metadata, WGS84 for data and OWL for linking data to external resources, such as DBPedia and Freebase. New terms have been defined only when they were not available in existing vocabularies.

This was a vertical tactical experimentation to gain momentum and engagement with the stakeholders in order to show that practical results can be obtained in a reasonable time and with reduced costs (with a minimal overhead for an on-going project). We retain that such an approach has been a success and it prepared and has opened the road for a larger transversal initiative.

ACKNOWLEDGMENTS

This work has been supported by the Autonomous Province of Trento, Italy. We are thankful to Roberto Bona, Isabella Bressan, Giulio De Petra, Marco Combetto, Luca Senter, Lorenzino Vaccari, Giuliano Carli, Fausto Giunchiglia, Maurizio Napolitano, Giovanni Tummarello, Michele Barbera, Piergiorgio Cipriano, Stefano Pezzi and the Trentino Open Data (TOD) group members for many fruitful discussions on the various aspects of releasing open government data covered in this report.

REFERENCES

1. European Parliament, “Directive 2007/2/EC establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)”, 2009.
2. T. Berners-Lee. Linked Data. Design Issues for the World Wide Web - W3C, http://www.w3.org/DesignIssues/LinkedData.html, 2006.
3. L. Manuel, V. Blazquez, B. Villazon-Terrazas, V. Saquicela, A. de Leon, O. Corcho, A. Gomez-Perez. Geolinked data and INSPIRE through an application case. In Proceedings of GIS, pages 446–449, 2010.
4. 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.
5. L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. Todd Williams, X. Li, J. Michaelis, A. Graves, J. 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, 9(3):325–333, 2011.
6. J. Euzenat, P. Shvaiko. Ontology Matching. Springer, Heidelberg (DE), 2007.
7. F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich. A faceted ontology for a semantic geo-catalogue. In Proceedings of ESWC, pages 169–182, 2011.
8. F.Giunchiglia, V. Maltese, F. Farazi, B. Dutta. GeoWordnet: A resource for geo-spatial applications. In Proceedings of ESWC, pages 121–136, 2010.
9. J. Goodwin, C. Dolbear, G. Hart. Geographical linked data: the administrative geography of Great Britain on the semantic web. Transaction in GIS, 12(1):19–30, 2009.
10. T. Heath, C. Bizer. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 2011.
11. F. J. Lopez-Pellicer, M. J. Silva, M.Chaves, J. F. Zarazaga-Soria, P. R. Muro-Medrano. Geo linked data. In Proceedings of DEXA, pages 495–502, 2010.
12. F. Maali, R. Cyganiak, V. Peristeras. A publishing pipeline for linked government data. In Proceedings of ESWC, pages 778–792, 2012.
13. D. Le Phuoc, A. Polleres, M. Hauswirth, G. Tummarello, C. Morbidoni. Rapid prototyping of semantic mash-ups through semantic web pipes. In Proceedings of WWW, pages 581–590, 2009.

14. A. Sala, C. Lin, H. Ho. Midas for government: Integration of government spending data on hadoop. In Proceedings of ICDE Workshops, pages 163–166, 2010.
15. A. Schellong, E. Stepanets. UncharteredWaters: The State of Open Data in Europe. CSC, Public Sector Study Series, 2011.
16. N. Shadbolt, K. O'Hara, M. Salvadores, H. Alani. eGovernment. In John Domingue, Dieter Fensel, and James Hendler, editors, Handbook of Semantic Web Technologies, pages 840–900. Springer, 2011.
17. P. Shvaiko, J. Euzenat. Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering, 2012, to appear.
18. P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi. A semantic geo-catalogue implementation for a regional SDI. In Proceedings of INSPIRE, 2010.
19. P. Smits, A. Friis-Christensen. Resource discovery in a European Spatial Data Infrastructure. IEEE Transactions on Knowledge and Data Engineering, 19(1):85–95, 2007.
20. N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, G. Tummarello. DING! Dataset Ranking using Formal Descriptions. In Proceedings of the Linked Data on the Web (LDOW) workshop at WWW, 2009.
21. E. Lynn Usery, D. Varanka. Design and Development of Linked Data for The National Map. The Semantic Web Journal, 2011.
22. L. Vaccari, P. Shvaiko, M. Marchese. A geo-service semantic integration in spatial data infrastructures. Int. Journal of Spatial Data Infrastructures Research, 4:24–51, 2009.
23. L. Vaccari, P. Shvaiko, J. Pane, P. Besana, M. Marchese. An evaluation of ontology matching in geo-service applications. GeoInformatica, 16(1):31–66, 2012.
24. G. Vickery. Review of recent studies on PSI re-use and related market developments. Information Economics, Paris, 2011.
25. J. Volz, C. Bizer, M. Gaedke, G. Kobilarov. Discovering and maintaining links on the web of data. In Proceedings of ISWC, pages 650–665, 2009.
26. J. Wielemaker, V. de Boer, A. Isaac, J. van Ossenbruggen, M. Hildebrand, G. Schreiber, S. Hennicke. Semantic workflow tool available. Europeana-Connect Deliv. 1.3.1, 2011.
27. D. Wood, editor. Linking Government Data. Springer, 2011.