Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

PEER REVIEWED PAPERS
STEPS TOWARD THE DIGITAL AGENDA: OPEN DATA TO OPEN KNOWLEDGE
Vol. 10, n.2, 2014
ISSN: 1826-6223 | eISSN: 1971-8829

OPEN DATA FROM EARTH OBSERVATION: FROM BIG DATA TO LINKED OPEN DATA, THROUGH INSPIRE

Massimo Zotti, Claudio La Mantia
Planetek Italia S.r.l.
zotti@planetek.it, lamantia@planetek.it

Keywords: Linked Open Data, Earth Observation, Big Data, INSPIRE

An efficient management of the increasing availability of Earth Observation and geographic data implies to respond to the paradigm of the 4V that usually applies to the problem of the Big Data: Volume - the sheer size of the “data at rest”, Velocity - the speed of new data arriving, Variety - the different manifold, and Veracity - trustworthiness and issues of provenance. Big data need to be quickly processed and analysed in conjunction with other data sources in order to express their real value in the construction of new knowledge. These processes are hastened by the advent of an increasing machine-to-machine communication. The automation of the data analysis requires standardized and linked data so that machines, without human intervention, can process them.

Standardization of geospatial data is mainly solved by the regulatory scenario dictated by the INSPIRE Directive[1]. The publication of spatial data

[1] http://inspire.ec.europa.eu/

for citations:
Zotti M., La Mantia C. (2014), *Open Data from Earth Observation: from Big Data to Linked Open Data, through INSPIRE,* Journal of e-Learning and Knowledge Society, v.10, n.2, 91-100. ISSN: 1826-6223, e-ISSN:1971-8829

as Linked Open Data may then leverage the reuse of common ontologies and vocabularies that allow the connection of geospatial data with other heterogeneous information. This way new scenarios and business opportunities may arise, as in the case of the real estate market mentioned in this article. This contribution aims to identify business opportunities, related to Linked Open Data and arising from the imminent availability of the Sentinel satellite data, with the European program Copernicus, for companies operating in the so-called downstream services of Earth observation.

1 Introduction

Copernicus[2] is the European program for the environmental monitoring that leverages the technology of Earth Observation to understand how our planet and its climate are changing, but also which is the impact of human activities on these changes, and how they will affect our daily lives.

Copernicus represents, if well used, an important tool to increase the security of European environmental policies for the sustainable development, but also to foster the development of a new economy based on the use of these data. A study by the European Commission has determined that in the next 15 years Copernicus will generate over 20,000 new direct jobs, compared to the current 5,000 employees in the sector. This will mainly happen in the "downstream", the sector where companies offer added value services on satellite remote sensing data. This sector is largely featured by SMEs, small and medium-sized companies that develop commercial applications based on data from Earth observation. Indeed, according to forecasts, the indirect benefit offered by the integration of these data with other information (cadastral data, meteorological, traffic information etc.) will have an even greater impact on overall employment in Europe, with more than 80,000 new jobs created by 2030.

As part of Copernicus, the Sentinel mission[3] is scheduled to launch by 2020: five families of satellites equipped with synthetic aperture RADAR and multi-spectral sensors for the monitoring of land, ice, oceans and atmosphere. There has recently been an extensive discussion about the licensing policy[4] to be applied to the Sentinels from the European Commission, with a strong pressure from the companies operating in the downstream sector of the remote sensing in favour of an open policy, and some rigidity by the providers of business data. The perspective nowadays is that the European states can count, thanks to Copernicus, on accurate, frequent and free of charge information, except for a few limitations.

The increasing availability of Earth Observation data must be looked in according to the paradigm of Big Data. Translated into the geospatial field, the Big Data paradigm is declined to the enormous volume of data now globally

[2] http://www.copernicus.eu/

[3] http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4

[4] http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Free_access_to_Copernicus_Sentinel_satellite_data

available and to the high frequency with which they are updated through new acquisitions, including remote sensing data.

One of the greatest challenges today is the ability to put in place quick processes for the extraction of information from multi-temporal series of Earth Observation data, which provide rapid and standardized quantitative information about the impact of natural emergencies (earthquakes, flooding) or the evolution of natural phenomena (desertification, land use, urban sprawl, etc.). Furthermore, the repetitiveness in the analysis of these data, allows the calculation of statistical indicators (for example the soil loss index) to support policies of territorial government and decision-making.

2 Added-value products from Earth Observation and Linked Open Data

Not only the national and international space agencies but even the commercial providers of satellite data are more and more providing EO data with licenses that allows the Value Added Providers (typically companies operating in the aforementioned downstream sector) to freely decide, in turn, which kind of license they want to apply to the so-called derivative products. Derivative products are the result of the elaboration of a satellite image through the creation of new geospatial information. Starting from the original satellite image, derivative products have different characteristics and properties from it and do not include the satellite dataset itself. Derivative products could be, for example, the vector information layers obtained from the classification or photointerpretation of a satellite image such as the existing buildings in a territory or the land use classification of that territory, which can be fully provided as Open Data if they not include the original satellite dataset. There are many examples of EO added-value products and the ability to distribute them as Open Data opens today new scenarios and business opportunities for companies who want to get through the traditional logic of thematic products supply.

Anyway, by comparing the high frequency with which the current (and future) satellite missions provide data and the slowness of the corresponding processing automation, there is a too long waiting for the creation of knowledge from these data. Very often, the creation of new knowledge implies the intersection of different information from different sources and, if this process is conducted in a traditional way by manual comparisons performed by human operators, the access to knowledge is inevitably delayed. Instead, the development of programmed workflows for the conversion of such data in Linked Open Data (LOD) allow an automatic access to them and their exploitation to become an automatic and independent process, not affected by human intervention.

The publication of data as LOD in fact is based on open and standard web technologies, aiming to provide information that can be read and understood by computers, in order to automatically connect and use data coming from different sources. The process starts from identification of the different geographical data sources; heterogeneous information has to be integrated using ETL (Extract, Tansform, Load) tools. Then an ontology has to be defined, which is specific to that type of data. It obviously preferable to use a common ontology: there are already a number of actions both at European level[5] and at national level in Italy[6], aiming to define a common vocabulary for ontologies that refers to the INSPIRE data models. The ontology can be generated using an editor like Protegè. Based on this ontology, the data can be converted in RDF format and published on a GEOSPARQL endpoint. At this phase, data can be queried using SPARQL language and visualized on a web map viewer. Furthermore, they can be linked with other existing data in the Web and relations can be inferred among them.

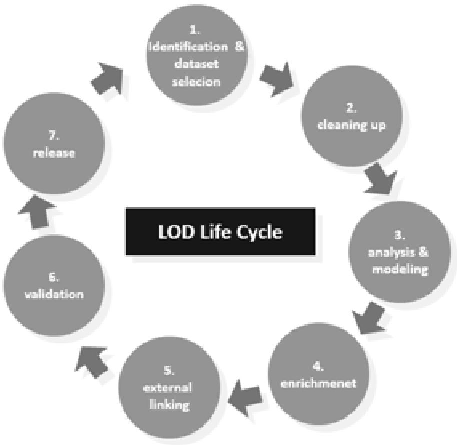

Fig. 1 – The Linked Open Data life cycle (source: AGID[7])

3 Linked Open Geographic Data

Italy has already gained an important experience on this issue, thanks to

[5] A. Perego, 2012: "Cross-domain interoperability for EU spatial data", http://www.w3.org/2012/06/pmod/pmod2012-jrc-andrea_perego.pdf; M. Lutz, A. Perego, M. Craglia, 2013: "Interoperability of (open) geospatial data: INSPIRE and beyond", http://www.w3.org/2013/04/odw/odw13_submission_58.pdf

[6] http://www.agid.gov.it/sites/default/files/documentazione_trasparenza/semanticinteroperabilitylod_en_3.pdf

[7] "Linee Guida per l'Interoperabilità Semantica attraverso i Linked Open Data", AGID -Agenzia per l'Italia Digitale, 2012 CdC-SPC-GdL6-InteroperabilitaSemOpenData_v2.0.doc

a system called GetLOD[8] for the publication of geographical Open Data as Linked Open Data from Spatial Data Infrastructures.

GetLOD is a software application created by Planetek Italia and SINERGIS, two leading companies in the field of geographic information in Italy, for the Emilia-Romagna regional administration, as part of the evolutionary development of the regional geographic infrastructure.

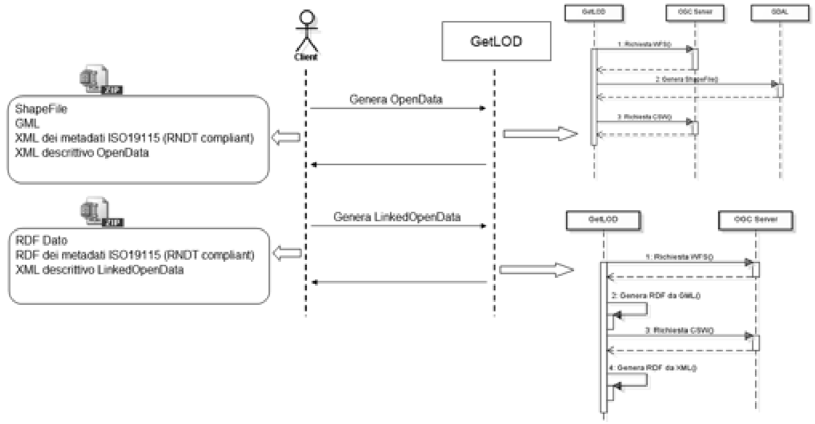

Fig. 2 – use of GetLOD to publish both Open Data and Linked Open Data

As shown in Figure 1, this application allows transforming data and web mapping services in open data, in accordance with the "five-star" classification by Tim Berners-Lee[9], meaning a standard format and structure that is directly usable by applications without manual intervention. Making geographic information as open data available, with particular attention to the RDF / XML, the use of data as Linked Open Data is enabled. At the same time it is enabled their re-use, the indexing in open data search engine and the integration with open data portals or with the Comprehensive Knowledge Archive Network[10] (CKAN), the catalog of free datasets and projects.

GetLOD is an open and reusable solution which can be integrated with any cartographic geoportal or Spatial Data Infrastructures (SDI) based on the interoperability standards defined by the Open Geospatial Consortium[11] (OGC®). The geographic data may be appraised through the provision both in RDF format and in other interchange formats (e.g. in Shape File format).

Planetek Italia and SINERGIS have realized the first components of the

[8] http://www.planetek.it/eng/products/all_products/getlod

[9] http://www.w3.org/DesignIssues/LinkedData.html

[10] http://ckan.org/

[11] http://www.opengeospatial.org/

solution within the evolutionary development of the geographical infrastructure of the Emilia-Romagna region. The goal of the project was to make available as Linked Open Data both the Data and Metadata handled by the SDI of the Emilia-Romagna region, in order to maintain the alignment and coordination between the different systems for open data publishing in the entire region. The regional SDI is in fact one node of the wider infrastructure of open data sharing in the region.

With reference to the data, the first classes of spatial objects published as Linked Open Data are the Addresses, Buildings, Geographical names and Administrative units: these geographical datasets represent crucial information coming from authoritative sources, to be used as reference data for interlinking external datasets from remote and heterogeneous sources in the cloud.

The definition of the ontologies, that describes the meaning of the data to be published, is preparatory to the actual publication of the RDF / XML data. In the case of Emilia-Romagna the conceptual modeling of objects to be published has not been defined from scratch, but rather borrowed from what was used in the Data Mart for the consultation of the regional Topographical Database.

Likewise, in the case of metadata, it has been necessary to define an ontology that describes the meaning of the ISO[12] 19115[13] metadata. These can be downloaded from the Regional Geoportal in XML format, according to the scheme defined by the ISO 19139[14] standard, and it is possible to map this schema to an OWL ontology, thus translating each metadata tab in a RDF / XML file based on this ontology.

The publication of data as LOD assumes significance when these data are linked to other existing data already published as LOD and are in turn potentially referable by others. This generally applies to any kind of data, but it is notably true for geographical data that are, by their nature, the basis for the correlation of information.

The transformation services for data and metadata are based on the use of the standards defined by the OGC® for Geographic webservices. The data to be published as Linked Open Data are extracted in RDF format using the standard OGC - WFS[15] (Web Feature Service) exposed by the regional Spatial Data Infrastructure for the access to geographic data. The extraction of geographic metadata in RDF format is instead performed using the standard OGC - CSW[16] (Catalog Service for the Web). This way the metadata can be connected to RDF

[12] http://www.iso.org/iso/home.html

[13] ISO 19115:2003 defines the schema required for describing geographic information and services. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

[14] ISO/TS 19139:2007 defines Geographic MetaData XML (gmd) encoding, http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

[15] http://www.opengeospatial.org/standards/wfs

[16] http://www.opengeospatial.org/standards/cat

(Linked Open Data) and to the shapefile data.

The dynamic generation of RDFs, starting from webservices exposed by the Geoportal, ensures the constant alignment with the generalist open data portal dati.emilia-romagna.it, which serves as the repository / indexing of generic open data and metadata (including RDF). The solution, finally, involves the use of free software to ensure the reusability of the system.

4 The contribution of INSPIRE

As already mentioned above, the reuse of common ontologies and vocabularies for geographical data may be effective only in standardized data models. The problem with the standardization of geospatial data is solved thanks to the respect of the rules defined by the INSPIRE Directive.

INSPIRE is a European Directive (2007/2/EC of 14 March 2007) that aims to achieve the interoperability of the spatial data infrastructures of the Member States, in order to support the governmental policies that have a direct or indirect impact on the European and global environment. Thanks to this important project the European Union, through its Spatial Data Infrastructure, must publish and share the geographical data, metadata and services of its Member States. It is very important to emphasize that the INSPIRE directive defines the rules for the sharing of data, metadata and services. These rules are described in detail in the Implementing Rules, that being issued by the European Commission are therefore binding legal acts, directly applicable in all Member States.

With reference to geographic data, INSPIRE dictates the rules of sharing for 34 types of thematic data (e.g. Geographical Names, Orthoimagery, Land Use, Species Distribution and so on), identified in Annexes I, II and III of the Directive.

Therefore today, those who work in the field of geomatics, know (or should know) that the geographic information related to the themes among listed, although obtained by heterogeneous processes, must comply to the standardized data model defined by INSPIRE and shared at European level.

As explained, the definition of common ontologies for INSPIRE data models is necessary to make European geospatial data automatically linkable in machine-to-machine applications.

It is worth noting that INSPIRE, from this point of view, represents a unique opportunity to overcome problems that are common in other areas of the Semantic Web, deriving from the extreme variety in sizes, in data structures and contents. On the contrary, the European Directive provides a clear and shared platform of technical and legal standards.

There are already several initiatives in Europe aiming to define common vocabularies for ontologies, with reference to the INSPIRE data models. It will be equally important to enhance the experiences already made at national level in the publication of Linked Open Data starting from geo-topographical database. The case mentioned above, made by the Emilia-Romagna region in Italy, is consistent with the aim of generating LOD starting from geographic databases that can be updated through the use of Earth observation data.

5 An example: Linked Open Data for the real estate market

Linked Open Data may create new business opportunities, and a clear example for geographic LOD may be seen when talking about the "Buildings" theme in the real estate market.

Thanks to open data, it is already possible to manually create mash-up applications that combine buying and selling data with information about the environment, transport, crime, etc. For example, you could geolocate on a map, in addition to houses, also Points Of Interest (POI) such as kindergartens, schools, museums, parks, public transport lines and so on. The cross analysis of these data may already provide some useful information to the buyer, thus representing already an opportunity to offer some added value for the real estate companies.

Furthermore, considering the temporal dimension, the analysis of multi-temporal images acquired by satellite offers the possibility to easily and automatically extract information about new buildings in a local context.

The importance of Linked Open Data is that the publication of this up-to-date information as LOD allows to automatically associate the information available in the LOD cloud to the new buildings. For example, the information about the builder, the zoning of the buildings and about any discounts available to buyers of the new properties, where this information is also published as Linked Open Data by the local administration.

The listings on the map, this way, would have an even higher value, and would be more complete and transparent. The buyers would acquire a greater knowledge of the area in which the property is situated, having more opportunities to make a reasoned choice – without the need for a human operator who cares to put together the data, and leaving it to the machines the burden of carrying out the analysis and comparison.

The amount of information available would significantly affect the transparency of the market, making the customers more informed and aware, and consequently decreasing the asymmetry between the seller and the buyer, thus reducing misconducts by agents of the sector.

Conclusion

This contribution aims to identify some business opportunities, deriving from Linked Open Data and arising from the imminent availability of the Sentinel Earth observation data from the European program Copernicus, for companies operating in the so-called downstream sector.

The open policy on Copernicus data, information and services will allow citizens, businesses, researchers and policy makers to integrate the environmental dimension in all their activities and decision-making procedures, ensuring greater transparency of policy choices.

The publication as Linked Open Data of added-value information extracted from satellite images can trigger a self-propelling process, which takes full advantage of the opportunities arising from the increasingly widespread availability of Earth observation data. The advantages are numerous and include a better access, for citizens and institutions, not only to geographic information but also to a real knowledge of the land and its dynamics, in support of Community, national and local policies.

The convergence of the activities of the European program Copernicus with the processes triggered by the INSPIRE Directive creates exciting opportunities for companies in the downstream sector of Earth observation that, in various ways, are positioned along the value added chain of EO data and services.

Thanks to the automation allowed by geographic Linked Open Data, the European industry can offer innovation in the different phases of the value adding process. For example, they can support the intelligent access to data in emergencies or access to catalogs for the best selection of scenes of interest. They can provide the image processing for the extraction of added-value information using rapid and standardized workflows to ensure standardized products and comparisons over time. They can support the harmonization and sharing of the data in interoperable formats and structures, in order to comply with European regulations, up to the publication as Linked Open Data, so that they can be enriched through the automatic connection to other data, creating new added value. The above mentioned are just a few of the possible benefits for the European SMEs that can bring to the creation of new jobs in Europe for professional with high-level skills.

REFERENCES

SpaceTec Partners, *Assessing the Economic Value of Copernicus: European Earth Observation and Copernicus Downstream Services Market Study* - Publishable Executive Summary – Final, URL: http://www.copernicus.eu/pages-principales/

library/study-reports/ (accessed on 16th April 2014).

EARSC, Open data study – Final report, URL: http://earsc.org/news/earsc-study-on-the-economic-benefits-of-a-free-and-open-data-policy-for-sentinel-satellite-data (accessed on 16th April 2014).

Karel Charvat, Tomas Mildorf, Jan Jezek, Karel Charvat Jr., Dmitrij Kozuch, Otakar Cerba, (2014), *Open Data for Real Estate Business*, W3C/OGC Workshop: Linking Geospatial Data, 5 - 6 March 2014, London.

Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman (2014), *Five Stars of Linked Data Vocabulary Use*, Semantic Web 0 1- 0. 1570-0844/14.

Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009), *Linked Data - The Story So Far*, International Journal on Semantic Web and Information Systems 5 (3): 1–22. ISSN 1552-6283.

Tim Berners-Lee (updated 2006/07/27), *Linked Data - Design Issues*. W3C. Retrieved 2010/12/18.

Fensel, Dieter; Facca, Federico Michele; Simperl, Elena; Ioan, Toma (2011), *Semantic Web Services*. Springer. p. 99. ISBN 3642191924.

Chris Bizer, Tom Heath, Kingsley Uyi Idehen, Tim Berners-Lee (2008), *Linked Data on the Web*. In Proceedings WWW2008, Beijing, China.

European Commission Joint Research Centre (2013), *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119*. URL: http://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf (accessed on 16th April 2014).

P. Denis, P. Jacques (2014), *OGC User Management Interfaces for Earth Observation Services*. http://www.opengis.net/doc/BP/EOUM/1.1 (accessed on 16th April 2014).

Claus Stadler, Jens Lehmann, Konrad Höffner, Sören Auer (2012), *LinkedGeoData: A core for a web of spatial open data*, Semantic web, Volume 3, Number 4 / 2012.

Mariana Damova, Atanas Kiryakov, Maurice Grinberg, Michael K. Bergman, Frédérick Giasson and Kiril Simov (2012), *Creation and Integration of Reference Ontologies for Efficient LOD Management*, Book DOI: 10.4018/978-1-4666-0188-8.ch007.

Durbha, S.S.; King, R.L. (2005), *Semantics-enabled framework for knowledge discovery from Earth observation data archives*, Geoscience and Remote Sensing, (Volume:43, Issue: 11), ISSN:0196-2892.