From Copernicus Big Data to Big Information and Big Knowledge: a Demo from the Copernicus App Lab Project

Konstantina Bereta
National and Kapodistrian University of Athens
konstantina.bereta@di.uoa.gr

Hervé Caumont
Terradue Srl
herve.caumont@terradue.com

Erwin Goor
VITO
erwin.goor@vito.be

Manolis Koubarakis
National and Kapodistrian University of Athens
koubarak@di.uoa.gr

Despina-Athanasia Pantazi
National and Kapodistrian University of Athens
dpantazi@di.uoa.gr

George Stamoulis
National and Kapodistrian University of Athens
gstam@di.uoa.gr

Sam Ubels
RAMANI B.V.
sam.ubels@ujuizi.com

Valentijn Venus
RAMANI B.V.
valentijn.venus@ujuizi.com

Firman Wahyudi
RAMANI B.V.
firman.wahyudi@ujuizi.com

ABSTRACT

Copernicus is the European program for monitoring the Earth. It consists of a set of complex systems that collect data from satellites and in-situ sensors, process this data and provide users with reliable and up-to-date information on a range of environmental and security issues. The data collected by Copernicus is made available freely following an open access policy. Information extracted from Copernicus data is disseminated to users through the Copernicus services which address six thematic areas: land, marine, atmosphere, climate, emergency and security. We present a demo from the Horizon 2020 Copernicus App Lab project which takes big data from the Copernicus land service, makes it available on the Web as linked geospatial data and interlinks it with other useful public data to aid the development of applications by developers that might not be Earth Observation experts. Our demo targets a scenario where we want to study the "greenness" of Paris.

KEYWORDS

satellite data, Linked Data, Copernicus

ACM Reference Format:
Konstantina Bereta, Hervé Caumont, Erwin Goor, Manolis Koubarakis, Despina-Athanasia Pantazi, George Stamoulis, Sam Ubels, Valentijn Venus, and Firman Wahyudi. 2018. From Copernicus Big Data to Big Information and Big Knowledge: a Demo from the Copernicus App Lab Project. In *Proceedings of International Conference on Information and Knowledge Management (CIKM2018).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Copernicus[1] is the European program for Earth observation. It collects data about our planet using a set of satellites (the Sentinel families) and contributing missions (existing commercial and public satellites). The first satellite (Sentinel-1A) was launched in 2014 and close to 20 satellites will be deployed by 2030. Copernicus also collects information from in-situ systems such as ground stations, which deliver data acquired by a multitude of sensors on the ground, at sea or in the air. The most recent study of the European Commission predicts that the cumulative economic value of Copernicus in the years 2008-2020 will be in the range of EUR 13.5 billion. The *Copernicus services* transform the wealth of satellite and in-situ Copernicus data into value-added products by processing and analysing the data. There are six Copernicus services covering the following thematic areas: atmosphere, marine, land, climate, emergency and security.

The case of Copernicus data is an interesting, real-world example of big data, as it gives rise to all the relevant big data challenges: volume, velocity, variety, veracity and value.

Data produced by the Copernicus services can be utilized in many applications with financial and environmental impact in areas such as emergency management, climate change, agriculture and security. This potential has not been fully realized up to now, because Copernicus services data "is hidden" in various archives operated by various European entities (e.g., the Flemish research institute VITO which participates in Copernicus App Lab). Therefore, a user that would like to develop an application needs to search in these archives, discover the needed data and integrate it in his application. The Copernicus App Lab project, whose results we will demonstrate in this demo, shows how to "break these silos open" by publishing their data as *linked data*, interlink it with other relevant data, and make it freely available online to enable the easy development of geospatial applications. Copernicus App Lab targets the variety and volume challenges of big Copernicus data.

CIKM2018, October 2018, Torino, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

Copernicus App Lab is a two-year European project funded by the Horizon 2020 program[2]. It is coordinated by the company AZO[3], and the technical partners are the National and Kapodistrian University of Athens[4], the companies Terradue[5] and RAMANI[6], and the Flemish research institute VITO[7].

The main innovations of the Copernicus App Lab project that we will demonstrate in our demo are the following. First, it enables users to access Copernicus data *on demand* using the popular OPeNDAP framework for accessing scientific data and GeoSPARQL APIs, instead of having to manually download, preprocess and integrate datasets coming from different sources and being in different formats. This is achieved by extending the ontology-based data access system Ontop-spatial [2] with the ability to support an OPeNDAP server as a data source. An ontology that describes the data conceptually and mappings from the data sources to the ontology are given as input. Mappings encode how the data provided by OPeNDAP can be translated into virtual RDF terms using the W3C mapping language R2RML [4]. Then, Copernicus data is accessible by posing GeoSPARQL queries; it can be retrieved from the data sources on demand and can be translated to RDF on-the-fly. This innovation targets the *variety* challenge of big data.

The second innovation of Copernicus App Lab targets the *volume* challenge of big data. The software components comprising the Copernicus App Lab platform (see next section) are available as Docker images and are deployed in the Terradue cloud platform, as cloud services, bringing the computation close to the data.

2 SOFTWARE COMPONENTS

Figure 1 shows the architecture of the Copernicus App Lab platform. In the following, we describe its main software components.

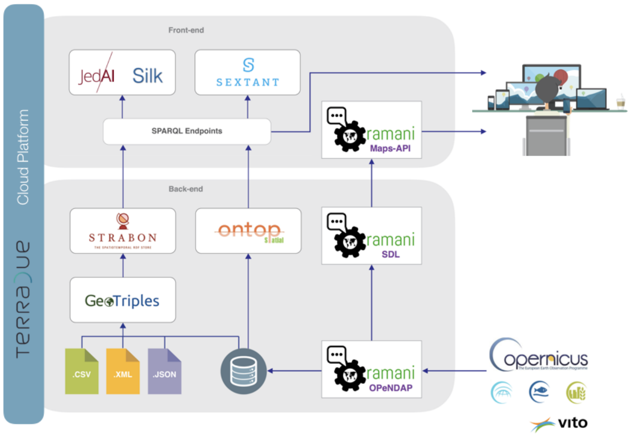


Figure 1: Copernicus App Lab architecture

2.1 The OPeNDAP framework and the streaming data library

Access to Copernicus data in Copernicus App Lab takes place through the popular *OPeNDAP framework* for accessing scientific data.[8] OPeNDAP provides a powerful data translation facility so that users do not need to know the detailed formats of data stored in servers, and can simply use a client that supports a model they are comfortable with.

The second core component of our software stack is the *streaming data library (SDL)*, developed by company RAMANI. The SDL interoperates with OPeNDAP in order to deliver *streams* out of loosely coupled Copernicus service providers. The SDL provides a set of tools that enable exploration, analysis and visualization of the data streams. Users can start by searching for datasets, view their metadata and select from a pool of analytic functions, such as temporal aggregation, to apply on selected variables. The output can then be returned through a variety of APIs (e.g., Android, Javascript, NodeJS).

In the version of the Copernicus App Lab software to be demonstrated, OPeNDAP and SDL are installed at the project partner VITO and provide direct access to the data archives of the Copernicus *global land service*.

2.2 Tools for linked geospatial data

In order to access Copernicus service products as linked data, we can download them, convert them into RDF, and store them in a triple store to make it accessible via a GeoSPARQL endpoint. While this workflow may be appropriate for certain products (e.g., the CORINE land cover dataset that is updated every 10 years), it introduces substantial overhead in cases where data gets updated frequently. In Copernicus App Lab, we tackle this problem by integrating the OPeNDAP API that RAMANI offers through the SDL, into the GeoSPARQL query engine Ontop-spatial. In this way, Ontop-spatial enables users to pose GeoSPARQL queries on top of OPeNDAP data sources without materialising any triples or tables.

The following linked geospatial data tools developed by the National and Kapodistrian University of Athens [7] are utilized in the project to make Copernicus services data available as linked data and interlink it with other data sources.

GeoTriples. GeoTriples is an open source tool that supports the automatic transformation of geospatial data from various geospatial formats (e.g., ESRI shapefiles, geospatial databases, XML and KML) into RDF using Semantic Web standards [9]. GeoTriples generates mappings that encode how data in their original format can be mapped to RDF terms. Both mapping languages R2RML and RML are supported to encode the mappings. R2RML [4] is a W3C standard language that was specifically designed for mapping relational data to RDF. RML [5] extends R2RML with the ability to support more file formats. GeoTriples uses the generated mappings to convert the input geospatial data into RDF.

JedAI and **Silk**. JedAI is a toolkit for entity resolution and its multi-core version has been shown to be scalable to very large datasets [11]. Silk is a well-known framework for interlinking heterogeneous data sources. Silk has been extended by us to support

[2] https://www.app-lab.eu/
[3] http://www.anwendungszentrum.de/
[4] http://kr.di.uoa.gr/
[5] https://www.terradue.com/
[6] https://ramani.ujuizi.com/
[7] https://remotesensing.vito.be/
[8] https://www.opendap.org/

the discovery of geospatial and temporal relationships among RDF resources [15].

Strabon. Strabon is a state-of-the-art spatiotemporal RDF store that supports storing spatiotemporal RDF data and evaluating spatiotemporal queries on it, using the query languages stSPARQL [8] and the OGC standard GeoSPARQL [13], that are geospatial extensions of the query language SPARQL [12]. According to benchmarks [6], Strabon is one of the most powerful, in terms of performance and functionality, geospatial RDF stores.

Ontop-spatial. Ontop-spatial is a geospatial ontology-based data access system able to integrate relational geospatial data sources using ontologies and mappings, and execute GeoSPARQL queries on top of them [2]. Ontop-spatial is a geospatial extension of the system Ontop [14]. In Copernicus App Lab, we extended Ontop-spatial with an OPeNDAP adapter which we implemented in the system MadIS [3], enabling it to support data sources that are not materialized locally in the system, but that can be fetched and transformed on-the-fly after firing a GeoSPARQL query to the system.

Sextant. Sextant is a Web-GIS tool for browsing and visualizing geospatial data available as KML, GML, image files and through WMS services, and communicating with various SPARQL endpoints to produce map layers out of the results of GeoSPARQL or stSPARQL queries [10]. All components of the Copernicus App Lab platform are available as dockers deployed in the Terradue cloud platform. All the above tools are open source and can be found at http://kr.di.uoa.gr/#systems.

3 PRESENTATION AND DEMONSTRATION

Our presentation will introduce the architecture of the Copernicus App Lab as managed on the Terradue cloud platform, and demonstrate the main technologies in a simple application. We will use a scenario from an environmental application whose aim is to study the evolution of green areas in Paris through time. We will showcase how this is achieved by using Earth Observation data in combination with other open geospatial data. More specifically, we will use the following datasets:

- A leaf area index dataset for Paris which is provided by the Copernicus global land service (by VITO) which will be accessed through the SDL and OPeNDAP. The *leaf area index (LAI)* is a dimensionless quantity that characterizes plant canopies. It is defined as the one-sided green leaf area per unit ground surface area [16]. This dataset is originally available in NetCDF format and it contains observations about the LAI values of areas at a given location (represented as a point) and time. Since all values contained in the NetCDF format are numeric, we aligned the spatial and temporal numeric values with OGC and W3C standards (i.e., all geometries are represented in WKT and timestamps are encoded in `xsd:dateTime` format). We also made this dataset accessible via a GeoSPARQL endpoint. We can access this data either on-the-fly using the tool Ontop-spatial, retrieving fresh observations from VITO as virtual RDF triples, or materialise these triples. Part of the materialized LAI dataset about Paris is available online on the popular DataHub platform[9].
- The most recent CORINE land cover (CLC) dataset which is available by the European Environment Agency (EEA) and contains information about the land cover of areas in Europe. Both the LAI and CLC datasets are offered by the land service of Copernicus. We created an INSPIRE-compliant ontology[10] for the CORINE land cover data, which conforms to the INSPIRE[11] requirements and belongs to the INSPIRE theme *Land Cover*[12]. Then, we transformed this dataset into RDF and published the resulting RDF dataset online[13].
- Urban Atlas. The Urban Atlas project[14] is providing pan-European comparable land use and land cover data for large urban zones with more than 100.000 inhabitants. This dataset is also provided by the EEA. We converted the Urban Atlas data into RDF. The RDF version of the dataset for every country is available online[15].
- The global administrative divisions dataset (GADM). This is an open dataset that contains information about the administrative divisions of countries worldwide and their subdivisions. We created an INSPIRE-compliant ontology [16] for the GADM data, which belongs to the INSPIRE theme *Administrative Units*.[17] Based on this ontology, we created the RDF version of the dataset for every country of the world at all levels of sub-division, and published it online[18].
- The OpenStreetMap (OSM) dataset that provides information about points of interest. We will use the OSM data about France in the form of shapefiles imported in a PostGIS database. This data is available as virtual RDF graphs using the OBDA system Ontop-spatial.

All datasets are available via GeoSPARQL endpoints and can be found on our website.[19]

The workflow of the demonstration scenario is the following:

(1) First, the user retrieves the most up-to-date LAI data using a GeoSPARQL query. This data is accessed using the OPeNDAP API, and it gets pre-processed and converted into virtual RDF triples on-the-fly using Ontop-spatial. The results are visualized in Sextant using a different colour for each LAI range. Using the timeline feature of Sextant, we can see the areas whose LAI changes through time (the colour of the points indicating these areas changes).

(2) Then, two more layers on the map will be created using the results of GeoSPARQL queries that retrieve CORINE land cover and Urban Atlas data about Paris that is available as linked data. These datasets can be compared with the LAI dataset, as we can see the land cover of areas that appear to be green according to the LAI dataset and especially the land cover of areas with LAI values that change through time.

(3) Next, we will do geospatial analytics combining the LAI, the CLC and the GADM dataset. For example, we can retrieve

[9] https://datahub.ckan.io/dataset/leaf-area-index-lai

[10] http://pyravlos-vm5.di.uoa.gr/corineLandCover.svg

[11] https://inspire.ec.europa.eu/

[12] http://inspire.ec.europa.eu/theme/lc

[13] https://datahub.ckan.io/dataset/corine-land-cover12

[14] https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-urban-atlas

[15] https://datahub.ckan.io/dataset/urban-atlas-12

[16] http://pyravlos-vm5.di.uoa.gr/gadmOntology.png

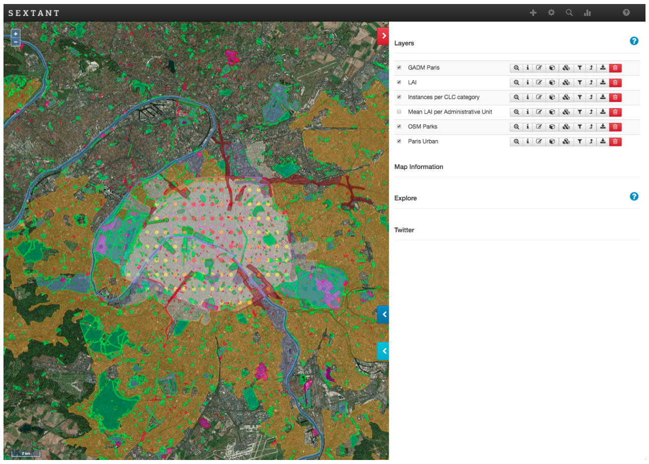
[17] http://inspire.ec.europa.eu/theme/au

[18] https://datahub.ckan.io/dataset/gadm

[19] http://kr.di.uoa.gr/#datasets

the average LAI value per administrative unit so that we can find out which is the“greener” administrative unit in Paris. Also, we can retrieve the most frequent CORINE land cover classes by aggregating the geometries of the areas in Paris that belong to the same CORINE land cover category and use different color per class to visualize this on the map.

(4) Last, the OSM dataset will be employed. We will then correlate the results of the CLC and LAI datasets about green areas with points of interest from OSM that are typically located in green areas (e.g., parks, forests, etc.).



☑ GADM Paris
☑ LAI
☑ Instances per CLC category
□ Mean LAI per Administrative Unit
☑ OSM Parks
☑ Paris Urban

Figure 2: Sextant map: “greenness” of Paris scenario

In Figure 2 we show a screenshot for the tool Sextant that provides a visualization of the data of our scenario. As we have already said, Sextant provides an interface for querying GeoSPARQL endpoints and presenting the results as layers on the world map. Each layer on the map of Figure 2 is the visualization of such a query. For example, the following query was used to retrieve the mean LAI area values for each administrative unit of Paris.

```
PREFIX lai:<http://dap.vgt.vito.be/thredds/dodsC/Copernicus/LAI/ontology/>
PREFIX corine:<http://geo.linkedopendata.gr/corine/ontology#>
PREFIX gadm:<http://geo.linkedopendata.gr/gadm/ontology#>
PREFIX geo:<http://www.opengis.net/ont/geosparql#>
PREFIX rdf:<http://www.w3.org/TR/rdf-schema/>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX strdf: <http://strdf.di.uoa.gr/ontology#>
SELECT  ?w1 ((AVG(?lai)) AS ?meanLAI) ?name
WHERE {
    ?s lai:LAI ?lai .
    ?s geo:hasGeometry ?geo .
    ?s lai:observationTime ?t .
    ?geo geo:asWKT ?w .
    ?adm a <http://geo.linkedopendata.gr/gadm/AdministrativeUnit> .
    ?adm gadm:hasName ?name .
    ?adm geo:hasGeometry ?geo1 .
    ?geo1 geo:asWKT ?w1 .
    ?adm gadm:belongsToAdm2 ?adm2 .
    ?adm2 gadm:hasName "Paris"^^<http://www.w3.org/2001/XMLSchema#string> .
    ?adm geo:hasGeometry ?geo2 .
    ?geo2 geo:asWKT ?w2 .
    ?adm strdf:intersects ?s.
}
GROUP BY ?w1 ?meanLAI ?name
```

Listing 1: GeoSPARQL query

The demo presented in this paper can be accessed on line.[20]

The software components of the Copernicus App Lab software stack have already undergone a first round of testing in an App Camp that took place in September 2017 and was organized by the European Space Agency[21] with the purpose of enabling developers with no expertise in Earth observation to develop interesting applications using state-of-the-art software technologies. More testing and experimentation is now underway by beta-testers, and a second App Camp is planned for September 2018.

ACKNOWLEDGEMENT

This work has been funded by the H2020 project Copernicus App Lab, Grant Agreement number 730124.

REFERENCES

[1] K. Bereta, H. Caumont, E. Goor, M. Koubarakis, D.-A. Pantazi, G. Stamoulis, S. Ubels, V. Venus, and F. Wahyudi. 2018. From Big Data to Big Information and Big Knowledge: the Copernicus App Lab Project. In *International Conference on Information and Knowledge Management (CIKM 2018), Case study/Industry paper. Submitted.*

[2] Konstantina Bereta and Manolis Koubarakis. 2016. Ontop of Geospatial Databases. In *Proceedings of the 15th International Semantic Web Conference.*

[3] Y. Chronis, Y. Foufoulas, V. Nikolopoulos, A. Papadopoulos, L. Stamatogiannakis, C. Svingos, and Y. E. Ioannidis. 2016. A Relational Approach to Complex Dataflows. In *Proceedings of the EDBT/ICDT Workshops 2016, Bordeaux, France.*

[4] Souripriya Das, Seema Sundara, and Richard Cyganiak. 2012. R2RML: RDB to RDF Mapping Language. W3C Rec.. (2012). Available from: http://www.w3.org/TR/r2rml/.

[5] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. RML: a generic language for integrated RDF mappings of heterogeneous data. In *LDOW.*

[6] George Garbis, Kostis Kyzirakos, and Manolis Koubarakis. 2013. Geographica: A Benchmark for Geospatial RDF Stores. In *the 12th International Semantic Web Conference, Sydney,Australia, October 21-25, 2013, Proceedings.* 343–359.

[7] M. Koubarakis, K. Bereta, G. Papadakis, D. Savva, and G. Stamoulis. 2017. Big, Linked Geospatial Data and Its Applications in Earth Observation. *IEEE Internet Computing* July/August (2017), 87–91.

[8] Manolis Koubarakis and Kostis Kyzirakos. 2010. Modeling and Querying Metadata in the Semantic Sensor Web: The Model stRDF and the Query Language stSPARQL. In *ESWC (LNCS)*, Vol. 6088. Springer, 425–439.

[9] Kostis Kyzirakos, Ioannis Vlachopoulos, Dimitrianos Savva, Stefan Manegold, and Manolis Koubarakis. 2014. GeoTriples: a Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, Riva del Garda, Italy, October 21, 2014.* 393–396.

[10] Charalampos Nikolaou, Kallirroi Dogani, Konstantina Bereta, George Garbis, Manos Karpathiotakis, Kostis Kyzirakos, and Manolis Koubarakis. 2015. Sextant: Visualizing time-evolving linked geospatial data. *Web Semantics: Science, Services and Agents on the World Wide Web* 35, 1 (2015).

[11] George Papadakis, Konstantina Bereta, Themis Palpanas, and Manolis Koubarakis. 2017. Multi-core Meta-blocking for Big Linked Data. In *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017.* 33–40.

[12] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. 2009. Semantics and Complexity of SPARQL. *ACM Trans. Database Syst.* 34, 3, Article 16 (Sept. 2009).

[13] Matthew Perry and John Herring. 2012. GeoSPARQL - A geographic query language for RDF data. Open Geospatial Consortium (OGC) Implementation Standard. (2012).

[14] Mariano Rodriguez-Muro and Martin Rezk. 2015. Efficient SPARQL-to-SQL with R2RML mappings. *Journal of Web Semantics* 33, 1 (2015).

[15] Panayiotis Smeros and Manolis Koubarakis. 2016. Discovering Spatial and Temporal Links among RDF Data. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with (WWW.*

[16] Wikipedia. 2018. Leaf Area Index —Wikipedia, The Free Encyclopedia. (2018). https://en.wikipedia.org/wiki/Leaf_area_index

[20] http://test.strabon.di.uoa.gr/SextantOL3/?mapid=m8s4kilcarub1mun_

[21] http://www.app-camp.eu/frascati/