A Data Science Pipeline for Big Linked Earth Observation Data

Manolis Koubarakis, Konstantina Bereta, Dimitris Bilidas, Despina-Athanasia Pantazi, and George Stamoulis

Abstract The science of Earth observation uses satellites and other sensors to monitor our planet, e.g., for mitigating the effects of climate change. Earth observation data collected by satellites is a paradigmatic case of big data. Due to programs such as Copernicus in Europe and Landsat in the United States, Earth observation data is open and free today. Users that want to develop an application using this data typically search within the relevant archives, discover the needed data, process it to extract information and knowledge and integrate this information and knowledge into their applications. In this chapter, we argue that if Earth observation data, information and knowledge are published on the Web using the linked data paradigm, then the data discovery, the information and knowledge discovery, the data integration and the development of applications become much easier. To demonstrate this, we present a data science pipeline that starts with data in a satellite archive and ends up with a complete application using this data. We show how to support the various stages of the data science pipeline using software that has been developed in various FP7 and Horizon 2020 projects. As a concrete example, our initial data comes from the Sentinel-2, Sentinel-3 and Sentinel-5P satellite archives, and they are used in developing the Green City use case.

Keywords Earth observation · Linked data · Big data · Knowledge graphs

1 Introduction

Earth observation (EO) is the science of using remote sensing technologies to monitor our planet including its land, its marine environment (seas, rivers and lakes) and its atmosphere. Satellite EO uses instruments mounted on satellite platforms to gather imaging data capturing the characteristics of our planet. These satellite

M. Koubarakis (✉) · K. Bereta · D. Bilidas · D.-A. Pantazi · G. Stamoulis
Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
e-mail: koubarak@di.uoa.gr

E. Curry et al. (eds.), *Technologies and Applications for Big Data Value*,
https://doi.org/10.1007/978-3-030-78307-5_19

images are then processed to extract information and knowledge that can be used in a variety of applications (e.g. in agriculture, insurance, emergency and security, or the study of climate change).

Lots of EO data are available to users at no charge today, due to the implementation of international programs such as Copernicus in Europe and Landsat in the United States. EO data is a paradigmatic case of big data bringing into play the well-known challenges of volume, velocity, variety, veracity and value. Regarding *volume*, according to the Copernicus Sentinel Data Access Annual Report of 2019 [14], the Sentinel satellites have produced *17.23 PiBs* of data from the beginning of operations until the end of 2019. Regarding *velocity*, the daily average volume of published data for the same satellites has been *18.47 TiBs* for November 2019. Regarding *variety*, EO data become useful only when analysed together with other sources of data (e.g. geospatial data or in situ data) and turned into information and knowledge. This information and knowledge is also big and similar big data challenges apply. For example, 1PB of Sentinel data may consist of about 750,000 datasets which, when processed, about 450TB of content information and knowledge (e.g. classes of objects detected) can be generated. Regarding *veracity*, EO data sources are of varying quality, and the same holds for the other data sources they are correlated with. Finally, the *economic value* of EO data is great. The Copernicus Market Report of 2019 [15] estimates that the overall investment of the European Union in the Copernicus program has been 8.2 billion Euros for the years 2008–2020. For the same period, the cumulated economic value of the program is estimated between 16.2 and 21.3 billion Euros.

Linked data is the data paradigm which studies how one can make RDF data (i.e. data that follow the Resource Description Framework[1]) available on the Web and interconnect it with other data with the aim of increasing its value. In the last few years, linked *geospatial* data has received attention as researchers and practitioners have started tapping the wealth of geospatial information available on the Web [19, 21]. As a result, the *linked open data (LOD) cloud* has been rapidly populated with geospatial data, some of it describing EO products (e.g. CORINE Land Cover and Urban Atlas published by project TELEIOS) [20]. The abundance of this data can prove useful to the new missions (e.g. the Sentinels) as a means to increase the usability of the millions of images and EO products that are expected to be produced by these missions.

However, big open EO data that are currently made available by programs such as Copernicus and Landsat are *not* easily accessible, as they are stored in different data silos (e.g. the Copernicus Open Access Hub[2]), and in most cases users have to access and combine data from these silos to get what they need. A solution to this problem would be to use Semantic Web technologies in order to publish the data contained in silos in RDF and provide semantic annotations and connections to them so that they can be easily accessible by the users. By this way, the value of the

[1] http://www.w3.org/TR/rdf-primer/.

[2] https://scihub.copernicus.eu/.

original data would be increased, encouraging the development of data processing applications with great environmental and processing value *even by users that are not EO experts but are proficient in Semantic Web technologies*.

The European project TELEIOS [20] was the first project internationally that has introduced the linked data paradigm to the EO domain, and developed prototype applications that are based on transforming EO products into RDF, and combining them with linked geospatial data. The ideas of TELEIOS were adopted and extended in the subsequent European projects LEO [8], MELODIES [7], BigDataEurope [2], Copernicus App Lab [3] and ExtremeEarth [18].

In this chapter, we present a data science pipeline that starts with data in a satellite archive and ends up with a complete application using this data. We show how to support the various stages of the data science pipeline using software developed by the above projects. As a concrete example, our initial data comes from the Sentinel-1 and Sentinel-5 satellite archives, and the developed application is the Green City use case we implemented in the context of project Copernicus App Lab[3].

The organization of the rest of the chapter is as follows. Section 2 introduces the Green City use case which serves the context for our application. Section 3 gives a high level of the data science pipeline and describes its various stages. Then, Sect. 4 describes how we have implemented the Green City use case using the linked geospatial data software developed in the projects mentioned above. Finally, Sect. 5 summarizes the paper.

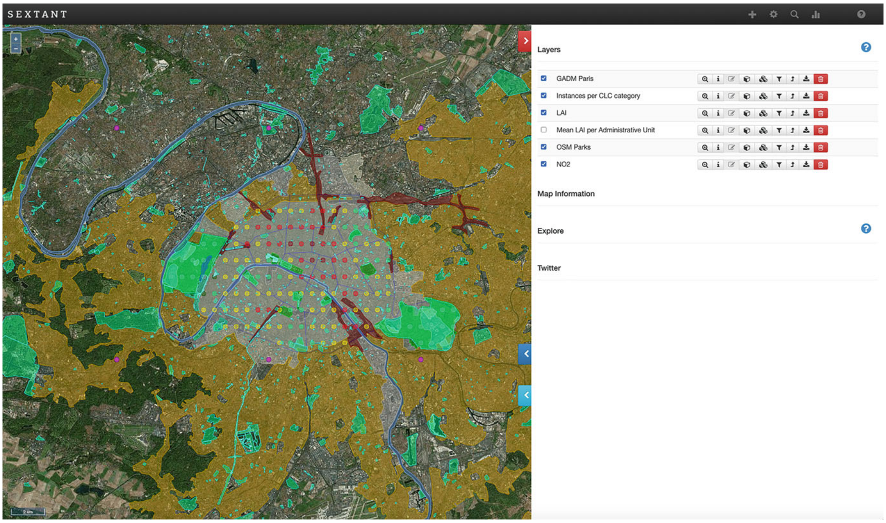
The chapter relates to the technical priority “Data Analytics” of the European Big Data Value Strategic Research and Innovation Agenda. It addresses the horizontal concern “Data Analytics” of the BDV Technical Reference Model. It addresses the vertical concern “Standards”.

The chapter relates to the “Knowledge and Learning” and “Systems, Methodologies, Hardware and Tools”, cross-sectorial technology enablers of the AI, Data and Robotics Strategic Research, Innovation and Deployment Agenda.

2 The Green City Use Case

Urban areas are the source of many of today’s environmental challenges – not surprisingly, since two out of three Europeans live in towns and cities. Local governments and authorities can provide the commitment and innovation needed to tackle and resolve many of these problems. The European Commission’s European Green Capital Award[3] (EGCA), recognizes and rewards local efforts to improve the environment, and thereby the economy and the quality of life in cities. The EGCA is given each year to a city, which is leading the way in environmentally friendly urban living. The award encourages cities to commit to ambitious goals for further environmental improvement.

[3] https://ec.europa.eu/environment/europeangreencapital/about-the-award/policy-guidance.



- ☑ GADM Paris
- ☑ Instances per CLC category
- ☑ LAI
- □ Mean LAI per Administrative Unit
- ☑ OSM Parks
- ☑ NO2

Fig. 1 A Green City map for Paris, France

Moreover, the Green City Accord[4] is a movement of European mayors committed to making cities cleaner and healthier. It aims to improve the quality of life of all Europeans and accelerate the implementation of relevant EU environmental laws. By signing the Accord, cities commit to addressing five areas of environmental management: air, water, nature and biodiversity, circular economy and waste, and noise.

In order to define though how "green" a city is, one must combine various sources of information that would allow us to measure and illustrate the greenness of each city in Europe. In the context of the Copernicus App Lab project, we demonstrated how one can interlink heterogeneous Earth Observation data sources and combine this information with other geospatial data using Linked Data technologies to produce Green City maps [3].

In Fig. 1 we show how to determine the greenness of Paris, France, by utilizing Earth Observation data, crowd-sourced data and Linked Data technologies. To produce this map, we combined air pollution data (NO_2 concentration) with indices that measure greenness (Leaf Area Index, OpenStreetMap Parks and CORINE Land Cover-related classes). All sources were spatially interlinked using the geometries of the administrative divisions of the city. Combining these diverse datasets using Linked Data technologies allows us to produce GeoSPARQL queries that can be visualized to construct such Green City maps for cities in Europe.

[4] https://ec.europa.eu/environment/topics/urban-environment/green-city-accord_en.

2.1 Data Sources

Sentinel data and Copernicus Services data that are currently available are not following the linked data paradigm. They are stored in different data silos so users might need to access and combine data from more than one source to satisfy their user needs. Utilizing Semantic Web and Linked Data technologies to make Copernicus Services data available as linked data increases their usability by EO scientists but also application developers that might not be EO experts. Moreover, the interlinking of Copernicus Services data with other relevant data sources (e.g. GIS data, data from the European data portal, etc.) increases the value of this data and encourages the development of applications with great environmental and financial value.

2.2 Copernicus Sentinel Data

For the Green City use case, the most relevant Earth Observation data come from the Land Monitoring service of Copernicus and air quality indices. To detect green areas within a city, we used the Leaf Area Index (Sentinel-3) and the CORINE land cover 2018 datasets (Sentinel-2 and Landsat-8). For air quality, we used the Nitrogen Dioxide index (Sentinel-5P).

The Leaf Area Index (LAI) is defined as half the total area of green elements of the canopy per unit horizontal ground area. The satellite-derived value corresponds to the total green LAI of all the canopy layers, including the understory which may represent a very significant contribution, particularly for forests. Practically, the LAI quantifies the thickness of the vegetation cover. LAI is recognized as an Essential Climate Variable by the Global Climate Observing System. The LAI dataset is provided by the Copernicus Global Land Service[5] and is distributed in Network Common Data Form version 4 (netCDF4) file format.

The CORINE Land Cover (CLC) inventory was initiated in 1985 (reference year 1990). Updates have been produced in 2000, 2006, 2012 and 2018. This vector-based dataset includes 44 land cover and land use classes. The time-series also includes a land-change layer, highlighting changes in land cover and land use. The high-resolution layers (HRL) are raster-based datasets which provide information about different land cover characteristics and is complementary to land cover mapping (e.g. CORINE) datasets. Five HRLs describe some of the main land cover characteristics: impervious (sealed) surfaces (e.g. roads and built up areas), forest areas, (semi-) natural grasslands, wetlands and permanent water bodies. The High-Resolution Image Mosaic is a seamless pan-European ortho-rectified raster mosaic based on satellite imagery covering 39 countries. The CLC dataset is provided by

[5] https://land.copernicus.eu/global/products/lai.

the Copernicus Pan-European component of the Land Monitoring Service[6] and is distributed in Shapefile format.

Nitrogen dioxide (NO_2) is a gaseous air pollutant composed of nitrogen and oxygen. NO_2 forms when fossil fuels such as coal, oil, gas or diesel are burned at high temperatures. NO_2 and other nitrogen oxides in the outdoor air contribute to particle pollution and to the chemical reactions that make ozone, thus it is one of six widespread air pollutants that have national air quality standards to limit them in the outdoor air. The NO_2 index is part of the Ozone Forecast dataset provided by the LOTOS-EUROS team, consisting of the Netherlands Organisation for Applied Scientific Research (TNO), the Environmental Assessment Agency of the Dutch National Institute for Public Health and the Environment (RIVM/MNP) and the Royal Netherlands Meteorological Institute (KNMI). The NO_2 dataset was distributed through the OPeNDAP protocol.

2.3 Other Geospatial Data

In addition to the above datasets, the Green City use case utilizes data from OpenStreetMap and the global administrative divisions dataset GADM.

OpenStreetMap (OSM) is a collaborative project to create a free editable map of the world. The geodata underlying the map is considered the primary output of the project. The creation and growth of OSM has been motivated by restrictions on use or availability of map data across much of the world, and the advent of inexpensive portable satellite navigation devices. The project has a geographically diverse user-base, due to emphasis of local knowledge and ground truth in the process of data collection. Many early contributors were cyclists who survey with and for bicyclists, charting cycleroutes and navigable trails. Others are GIS professionals who contribute data with Esri tools. In this manner, OSM is an open and free map of the whole world constructed by volunteers. It is available in vector format as shapefiles from the German company Geofabrik.[7] For our use case, information about parks has been taken from this dataset.

The Global Administrative Areas (GADM) dataset is a high-resolution database of country administrative areas, with a goal of “all countries, at all levels, at any time period”.[8] It is available in vector format as a shapefile, a geopackage (for SQLlite3), a format for use with the programming language R, and KMZ (compressed KML). GADM allows us to use the administrative boundaries of cities and spatially interlink it with all the information we have from the other datasets.

[6] https://land.copernicus.eu/pan-european/corine-land-cover.

[7] http://download.geofabrik.de/.

[8] https://gadm.org/.

3 The Data Science Pipeline

Developing a methodology and related software tools that support the complete life cycle of linked open EO data has been studied by our group in project LEO [21] following similar work for linked data, for example by project LOD2 and others [1, 27]. Capturing the life cycle of open EO data and the associated entities, roles and processes of public bodies and making available this data was the first step in achieving LEO's main objective of bringing the linked data paradigm to EO data centres, and re-engineering the life cycle of open EO data based on this paradigm. In this chapter we continue this work by presenting a data science pipeline for big linked EO data and we apply it to the development of the Green City use case presented in the previous section.

The life of EO data starts with its generation in the ground segment of a satellite mission. The management of this so-called payload data is an important activity of the ground segments of satellite missions. Figure 2 gives a high-level view of the data science pipeline for big linked EO data as we envision it in our work. Each phase of the pipeline and its associated software tools is discussed in more detail below.

3.1 *Ingestion, Processing, Cataloguing and Archiving*

Raw data, often from multiple satellite missions, is ingested, processed, catalogued and archived. Processing results in the creation of various standard products (Level 1, 2, etc., in EO jargon; raw data is Level 0) together with extensive metadata describing them.

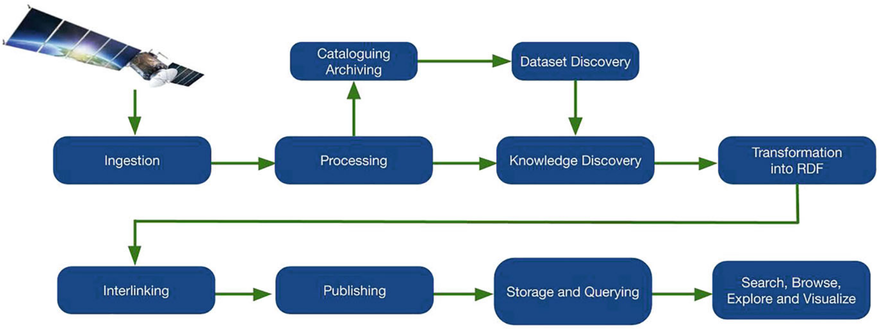


Fig. 2 The data science pipeline for big, linked EO data

3.2 Dataset Discovery

Once data become available in an archive or on the Web, they can be accessed by proprietary systems, traditional search engines or the Dataset Search service offered by Google.[9]

For example, the Copernicus Open Access Hub currently stores products from Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5P missions[10] and offers a menu/map interface for searching for relevant data by date/time, area of interest, mission, satellite platform, etc. Similar interfaces are offered by other EO data centres hosting satellite data such as NASA[11] and the German Aerospace Center DLR.[12]

An interesting recent development in the area of dataset search is the development of the service Dataset Search by Google. This service crawls the Web retrieving metadata of datasets annotated using Schema.org vocabularies[13] following the guidelines of Google researchers.[14] Schema.org was originally founded by Google, Microsoft, Yahoo! and Yandex, and it has evolved into a community activity developing vocabularies for annotating Web resources by an open community process. Schema.org provides a unique structured data markup schema to annotate a Web page with variety of tags that can be added to HTML pages as JSON-LD, Microdata or RDFa markup. This markup allows search engines to index Web pages more effectively.

Dataset Search also offers a keyword-based search interface for discovering these datasets. For example, one can search for "CORINE land cover Copernicus App Lab" to discover the CORINE land cover dataset in linked data form published on datahub.io by our project Copernicus App Lab.[15] CORINE land cover is a dataset published by the European Environment Agency describing the land cover/land use of geographical areas in 39 European countries.

3.3 Knowledge Discovery

In the knowledge discovery frameworks developed in project TELEIOS [12, 13], traditional raw data processing has been augmented with *content extraction* methods that deal with the specificities of satellite images and derive image descriptors

[9] https://datasetsearch.research.google.com/.

[10] https://scihub.copernicus.eu/.

[11] https://search.earthdata.nasa.gov/.

[12] https://eoweb.dlr.de/egp/.

[13] https://schema.org/.

[14] https://support.google.com/webmasters/thread/1960710.

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

(e.g. texture features, spectral characteristics of the image). Knowledge discovery techniques combine image descriptors, image metadata and auxiliary data (e.g. GIS data) to determine concepts from a domain ontology (e.g. park, forest, lake, etc.) that characterize the content of an image.

Hierarchies of domain concepts are formalized using *ontologies* encoded in the Web Ontology Language OWL2 and are used to annotate standard products. Annotations are expressed in RDF and its geospatial extension stRDF/GeoSPARQL [23, 30] and are made available as linked data so that they can be easily combined with other publicly available linked data sources (e.g. GeoNames, OpenStreetMap, DBpedia) to allow for the expression of rich user queries.

3.4 Transformation into RDF

This phase transforms vector or raster EO data from their standard formats (e.g. ESRI Shapefile or NetCDF) into RDF.

In FP7 project LEO we developed the tool GeoTriples for transforming EO data and geospatial data into RDF [26]. GeoTriples is able to deal with *vector data and their metadata* and to support natively many popular geospatial data formats (e.g. shapefiles, spatially enabled DBMS, KML, GeoJSON, etc.) The mapping generator of GeoTriples employs the mapping languages R2RML [10] and RML [11] to create mappings that dictate the method of conversion of the raw data into RDF.

R2RML is a language for expressing mappings from relational data to RDF terms, and RML is a more general language for expressing mappings from files of different formats (e.g. CSV, XML, etc.) to RDF. The mappings are enriched with subject and predicate object maps in order to properly deal with the specifics of geospatial data and represent it using an appropriate ontology.

GeoTriples is an open-source tool[16] that is distributed freely according to the Mozilla Public License v2.0.

3.5 Interlinking

This is a very important phase in the linked EO data life cycle since a lot of the value of linked data comes through connecting seemingly disparate data sources to each other.

Starting in our project LEO, we have worked on interlinking of open EO data by discovering geospatial or temporal semantic links. For example, in linked EO datasets, it is often useful to discover links involving topological relationships, for example `A geo:sfContains F`, where `A` is the area covered by a remotely

[16] http://geotriples.di.uoa.gr/.

sensed multispectral image `I`, `F` is a geographical feature of interest (field, lake, city, etc.) and `geo:sfContains` is a topological relationship from the topology vocabulary extension of GeoSPARQL. The existence of this link might indicate that `I` is an appropriate image for studying certain properties of `F`.

In LEO we have dealt with these issues by extending the well-known link discovery tool Silk in order to be able to discover precise geospatial and temporal links among RDF data published using the tool GeoTriples. The extension of Silk that we developed is now included in the main version. Since then other tools that carry out the same task more efficiently have been developed, for example Radon [32]. A recent comparison of geospatial interlinking systems is presented in [31].

3.6 Publishing

This phase makes linked EO data publicly available in the LOD cloud or in open data platforms such as datahub.io using well-known data repository technologies such as CKAN. In this way, others can discover and share this data and duplication of effort is avoided.

3.7 Storage and Querying

This phase deals with storing all relevant EO data and metadata on persistent storage so they can be readily available for querying in subsequent phases.

In our projects we have used our own spatiotemporal RDF store Strabon[17] which was developed especially for this purpose [24]. Strabon supports the data model stRDF and the query language stSPARQL developed by our group.

stRDF is an extension of RDF that allows the representation of geospatial data that changes over time [5, 25]. stRDF is accompanied by stSPARQL, an extension of the query language SPARQL 1.1 for querying and updating stRDF data. stRDF and stSPARQL use OGC standards (WKT and GML) for the representation of temporal and geospatial data.

Strabon extends the well-known open-source RDF store Sesame 2.6.3 and uses PostgreSQL or MonetDB as the backend spatially enabled DBMS. As shown by our experiments in [5, 16, 17, 25], Strabon is currently the most functional and performant geospatial and temporal RDF store available.

Strabon also supports the Open Geospatial Consortium (OGC) standard GeoSPARQL [30] for querying geospatial data encoded in RDF. stSPARQL and GeoSPARQL are very similar languages although they have been developed

[17] http://strabon.di.uoa.gr.

independently. Strictly speaking, if we omit aggregate geospatial functions from stSPARQL, the geospatial component of GeoSPARQL offers more expressive power than the corresponding component of stSPARQL. However, GeoSPARQL does not support a temporal dimension to capture the valid time of triples as stSPARQL does.

In our work stRDF has been used to represent satellite image metadata (e.g. time of acquisition, geographical coverage), knowledge extracted from satellite images (e.g. a certain area is a park) and auxiliary geospatial data sets encoded as linked data. One can then use stSPARQL to express in a single query an information request such as the following: "Find an image taken by a Meteosat second generation satellite on August 25, 2007, which covers the area of Peloponnese and contains hotspots corresponding to forest fires located within 2 km from a major archaeological site." Encoding this information request today in a typical interface to an EO data archive such as the ones discussed above is impossible, because domain-specific concepts such as "forest fires" are not included in the archive metadata, thus they cannot be used as search criteria.

With the techniques of knowledge discovery developed in our projects, we can characterize satellite image regions with concepts from appropriate ontologies (e.g. landcover ontologies with concepts such as waterbody, lake and forest, or environmental monitoring ontologies with concepts such as forest fires and flood) [13, 22]. These concepts are encoded in OWL2 ontologies and are used to annotate EO products. Thus, we attempt to close the semantic gap that exists between user requests and searchable information available explicitly in the archive.

But even if semantic information was included in the archived annotations, one would need to join it with information obtained from auxiliary data sources to answer the above query. Although such open sources of data are available to EO data centres, they are not used currently to support sophisticated ways of end-user querying in Web interfaces such as the ones discussed above under "Dataset Discovery". In our work, we have assumed that auxiliary data sources, especially geospatial ones, are encoded in stRDF and are available as linked geospatial data, thus stSPARQL can easily be used to express information requests such as the above.

In some applications it might not be a good idea to transform existing geospatial data into RDF and then store it in a triple store such as Strabon (e.g. when such data get frequently updated and/or are very large or when the data owners choose not to do so). For this case, we have developed the system Ontop-spatial,[18] which is a geospatial extension of the ontology-based data access (OBDA) system Ontop [9]. Ontop performs on-the-fly SPARQL-to-SQL translation on top of relational databases using ontologies and mappings. Ontop-spatial extends Ontop by enabling on-the-fly GeoSPARQL-to-SQL translation on top of geospatial databases [4, 6]. Ontop-spatial allows geospatial data to remain in their original

[18] http://ontop-spatial.di.uoa.gr.

databases (e.g. PostGIS, SpatiaLite, Oracle Spatial and Graph) and enables them to be queried effectively and efficiently using GeoSPARQL and the OBDA paradigm.

3.8 Search/Browse/Explore/Visualize

This phase enables users to find and explore the data they need and start developing interesting applications.

In FP7 project LEO, we redesigned the tool Sextant [28] for such purposes and also developed a mobile version that is distributed as an APK file for Android OS. The new version of Sextant is a web-based and mobile-ready application for exploring, interacting and visualizing time-evolving linked geospatial data.

Sextant was designed as an open-source application[19] that is flexible, portable and interoperable with other GIS tools. This allows us to use it as a core building block for creating new web or mobile applications, utilizing the provided features. The core feature of Sextant is the ability to create thematic maps by combining geospatial and temporal information that exists in a number of heterogeneous data sources ranging from standard SPARQL endpoints to SPARQL endpoints following the standard GeoSPARQL defined by the OGC, or well-adopted geospatial file formats, like KML, GML and GeoTIFF. In this manner we provide functionality to domain experts from different fields in creating thematic maps, which emphasize spatial variation of one or a small number of geographic distributions. Each thematic map is represented using a map ontology that assists on modelling these maps in RDF and allows for easy sharing, editing and search mechanisms over existing maps.

4 Implementing the Green City Use Case Using Linked Geospatial Data Software

In this section we present the implementation of the Green City use case using the pipeline of the previous section and the relevant software for each stage of the pipeline.

4.1 Ingestion

In Green City use case, access to Copernicus data and information was achieved in two ways: (1) by downloading the data via the Copernicus Open Access Hub or

[19] http://sextant.di.uoa.gr/.

the Websites of individual Copernicus services, and (2) via the popular OPeNDAP framework[20] for accessing scientific data.

4.2 Dataset Discovery

The Copernicus Open Access Hub[21] offers access to Sentinel data, using a simple graphical interface that enables users to specify the extent of the geographical area one is interested in. In the Green City use case though, we mainly used data from the Land Monitoring service that processes Copernicus data and produces higher-level products that are of importance in the corresponding thematic area. To detect green areas within the cities, we used the Leaf Area Index dataset, produced by Sentinel-3 data, and the CORINE land cover dataset for 2018, produced by Sentinel-2 and Landsat-8 (gap filling) data. For air quality, we used the Nitrogen Dioxide index, produced by Sentinel-5p data. Moreover, we used data from OpenStreetMap (OSM) and the Database of Global Administrative Areas (GADM).

4.3 Knowledge Discovery

Although in the Green City use case this step of the pipeline was not needed, it is a very crucial step that allows us to discover knowledge hidden in the EO images and use ontologies to describe this knowledge. Such techniques were used in the context of the projects TELEIOS and ExtremeEarth by our group in collaboration with Remote Sensing scientists. In TELEIOS, colleagues from the National Observatory of Athens developed algorithms to detect fires in SEVIRI images in the context of a fire monitoring application [20]. In ExtremeEarth, colleagues from the University of Trento perform the accurate crop type mapping needed the Food Security use case, using a deep learning architecture for Sentinel-2 images [29].

4.4 Transformation into RDF

In this stage of the pipeline, the outputs of the previous two stages are transformed into RDF, so that they can be combined with other interesting linked geospatial data. In the Green City use case, RDF is used to represent Earth Observation data produced by the Copernicus Land Monitoring service, air quality indices,

[20] https://www.opendap.org/.

[21] https://scihub.copernicus.eu/.

OpenStreetMap data and data from the database of Global Administrative Areas, as described in Sect. 4.2.

To transform the mentioned data into RDF, we developed INSPIRE-compliant ontologies. In the process of constructing ontologies to model Copernicus and other geospatial data, our aim is to provide standard-compliant, reusable and extensible ontologies. In this direction, we opted to follow vocabularies that have been defined in well-established standards, such as the INSPIRE directives and the OGC.

The INSPIRE directive aims to create an interoperable spatial data infrastructure for the European Union, to enable the sharing of spatial information among public sector organizations and better facilitate public access to spatial information across Europe.[22] INSPIRE-compliant ontologies are ontologies which conform to the INSPIRE requirements and recommendations. Our initial approach was to reuse existing INSPIRE-compliant ontologies, but since these efforts are not as close to the INSPIRE specifications as we would like to, we decided to construct our own INSPIRE-compliant versions, following the data specifications as closely as possible. Our aim is to reuse these ontologies for other datasets that belong to the same INSPIRE themes and also publish them so that others can reuse these ontologies for their geospatial datasets as well.

The ontologies we constructed for the Green City use case are the following:

- The ontology for the global database of *Leaf Area Index (LAI)*, as shown in the link: http://pyravlos-vm5.di.uoa.gr/laiOntology.png.
- The *CORINE Land Cover (CLC) ontology*, included in the link http://pyravlos-vm5.di.uoa.gr/corineLandCover.svg, shows the ontology constructed for the CLC dataset. The ontology is a specialization of the general ontology that we constructed to model the respective Land Cover theme of INSPIRE so that we have the first INSPIRE-compliant ontology.
- The ontology for the *Ozone Forecast* dataset, including the NO_2 index, as described in this link:
  http://pyravlos-vm5.di.uoa.gr/atmosphereTimeSeriesOntology.png.
- The *OpenStreetMap (OSM)* ontology, as shown in this figure: http://sites.pyravlos.di.uoa.gr/dragonOSM.svg.
- The ontology for the *Database of Global Administrative Areas (GADM)*, included in this link: http://pyravlos-vm5.di.uoa.gr/gadmOntology.png.

Figure 3 provides the ontology we constructed for the GADM dataset. To construct this ontology, we extended the GeoSPARQL ontology (namespaces `sf` and `geo`). For the class and properties that we introduced we use the prefix `gadm`.[23] The GADM ontology can be used so that a GADM dataset[24] can be either converted into RDF or queried on-the-fly.

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

[23] The corresponding namespace is: http://www.app-lab.eu/gadm/.

[24] https://gadm.org/data.html.

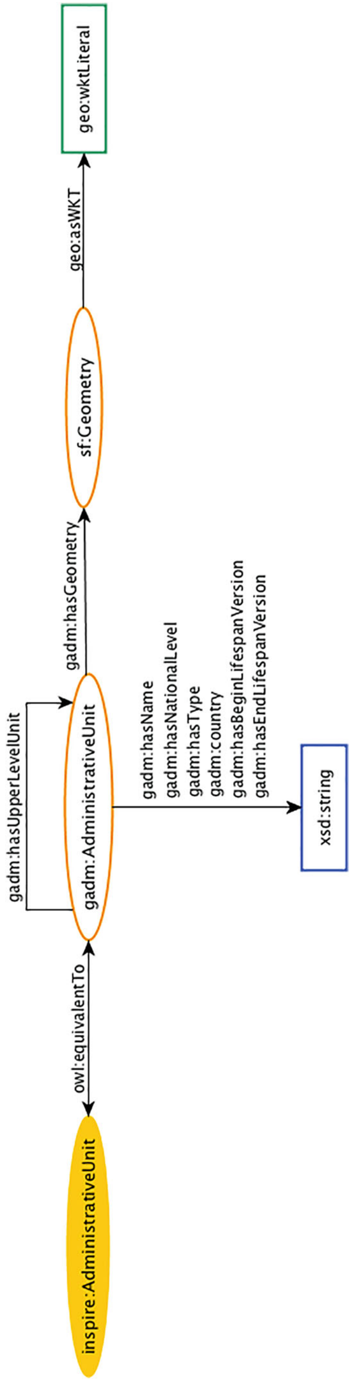


Fig. 3 The GADM ontology

For the transformation of the aforementioned datasets, we use the tool GeoTriples that automatically produces RDF graphs according to a given ontology. Shapefiles along with corresponding ontologies are provided as input to GeoTriples, which automatically creates R2RML or RML mappings that dictate the method of conversion of data into the RDF data model. Spatial information is mapped into RDF according to the GeoSPARQL vocabulary. Since GeoTriples does not support NetCDF files as input, in the case of the LAI dataset, the translation into RDF was done by writing a custom Python script.

It is very important to make the above datasets available on the Web as linked data, in order to increase their use, as, in this way, they can be made “interoperable” and more valuable when they are linked together. To achieve this goal, we followed Google Dataset Search guidelines and annotated all the datasets of the Green City use case by using the markup format JSON-LD. All these datasets can be searched and found using Google Dataset Search.

4.5 Storage/Querying

For storage and querying, we used the tools Strabon and Ontop-spatial. The spatiotemporal RDF store Strabon and the query languages stSPARQL and GeoSPARQL are used for storage and querying linked geospatial data originating from transforming EO products into RDF.

Strabon was utilized to create the SPARQL endpoints for the GADM, CLC 2018 and OSM parks data sources that are originally distributed in vector formats and are not updated frequently. For example, assuming appropriate PREFIX definitions, the GeoSPARQL query shown in Listing 1 retrieves how many CORINE areas in Paris belong to every land use category and projects the union of the geometries of these areas per category.

Listing 1 CORINE areas in Paris for every land cover category and their geometries

```
SELECT DISTINCT ?landUse (strdf:union(?w3) as ?geo) (count(?c) as
    ?instances)
WHERE{
  ?adm rdf:type gadm:AdministrativeUnit .
  ?adm gadm:hasName ?name .
  ?adm gadm:belongsToAdm2 ?adm2 .
  ?adm2 gadm:hasName
        "Paris"^^<http://www.w3.org/2001/XMLSchema#string> .
  ?adm geo:hasGeometry ?geo2 .
  ?geo2 geo:asWKT ?w2 .
  ?c corine:hasLandUse ?landUse .
  ?c geo:hasGeometry ?geo3 .
  ?geo3 geo:asWKT ?w3 .
  FILTER(geof:sfIntersects(?w2,?w3))}
GROUP BY ?landUse
```

For the rest of the data sources (LAI and NO_2) that are updated regularly and are distributed in raster formats, we chose to use Ontop-spatial. This solution does not require the transformation of the source data into RDF and allows us to create virtual RDF graphs on top of geospatial databases and data delivered through the OPeNDAP protocol, so they can be readily available for querying. In this case, the developer has to write R2RML mappings expressing the correspondence between a data source and classes/properties in the corresponding ontology. An example of such a mapping is provided in Listing 2, in the native mapping language of Ontop-spatial which is less verbose than R2RML.

Listing 2 Example of mappings

```
mappingId opendap_mapping
target lai:{id} rdf:type lai:Observation .
       lai:{id} lai:lai {LAI}^^xsd:float;
            time:hasTime {ts}^^xsd:dateTime .
       lai:{id} geo:hasGeometry _:g .
       _:g geo:asWKT {loc}^^geo:wktLiteral .
source SELECT id, LAI, ts, loc
       FROM (ordered opendap
       url:https://analytics.ramani.ujuizi.com/
       thredds/dodsC/Copernicus-Land-timeseries-
       global-LAI%29/readdods/LAI/, 10)
       WHERE LAI > 0
```

In the example mappings shown in Listing 2, the `source` is the LAI dataset discussed above, while the `target` part of the mapping encodes how the relational data is mapped into RDF terms. Given the mapping provided above, we can pose the GeoSPARQL query provided in Listing 3 to retrieve the LAI values and the geometries of the corresponding areas.

Listing 3 Query retrieving LAI values and locations

```
SELECT DISTINCT ?s ?wkt ?lai
WHERE { ?s lai:hasLai ?lai .
        ?s geo:hasGeometry ?g .
        ?g geo:asWKT ?wkt }
```

4.6 Publishing

Some of the RDF datasets that are used in the Green City use case have been published in the datahub https://datahub.ckan.io/organization/app-lab.

4.7 Interlinking

In the Green City use case, combining information from different geospatial sources was crucial, as we needed to spatially interlink the administrative divisions of a city with the EO data and OSM parks. We address this issue by employing the geospatial and temporal component of the framework Silk,[25] which is a component that enables users to discover a wide variety of spatial and temporal relations, such as intersects, contains, before, and during, between different sources of data.

To retrieve features for which a spatial relation holds (e.g., intersection and containment), we ask Silk to search for these relations between two RDF data sources given the relations' definitions. The outcome contains all of the entities for which the relations hold. For example, to interlink the CLC and the GADM datasets, a CLC class that intersects an administrative division is interlinked with it with the property *geo:sfIntersects*. The discovered relations are then materialized in the RDF store, resulting in a more semantically informative dataset.

Interlinking with topological and temporal relations can be used to considerably decrease the query response time by replacing the spatial and temporal functions with the respective bindings. For example, we can pose a SPARQL query by replacing the function *geof:sfIntersects* with the triple pattern *?clc geo:sfIntersects ?ad*, as the geospatial features for which the relation *geo:sfIntersects* holds have already been discovered, and the evaluation engine would simply have to retrieve the respective bindings instead of calculating the spatial filter.

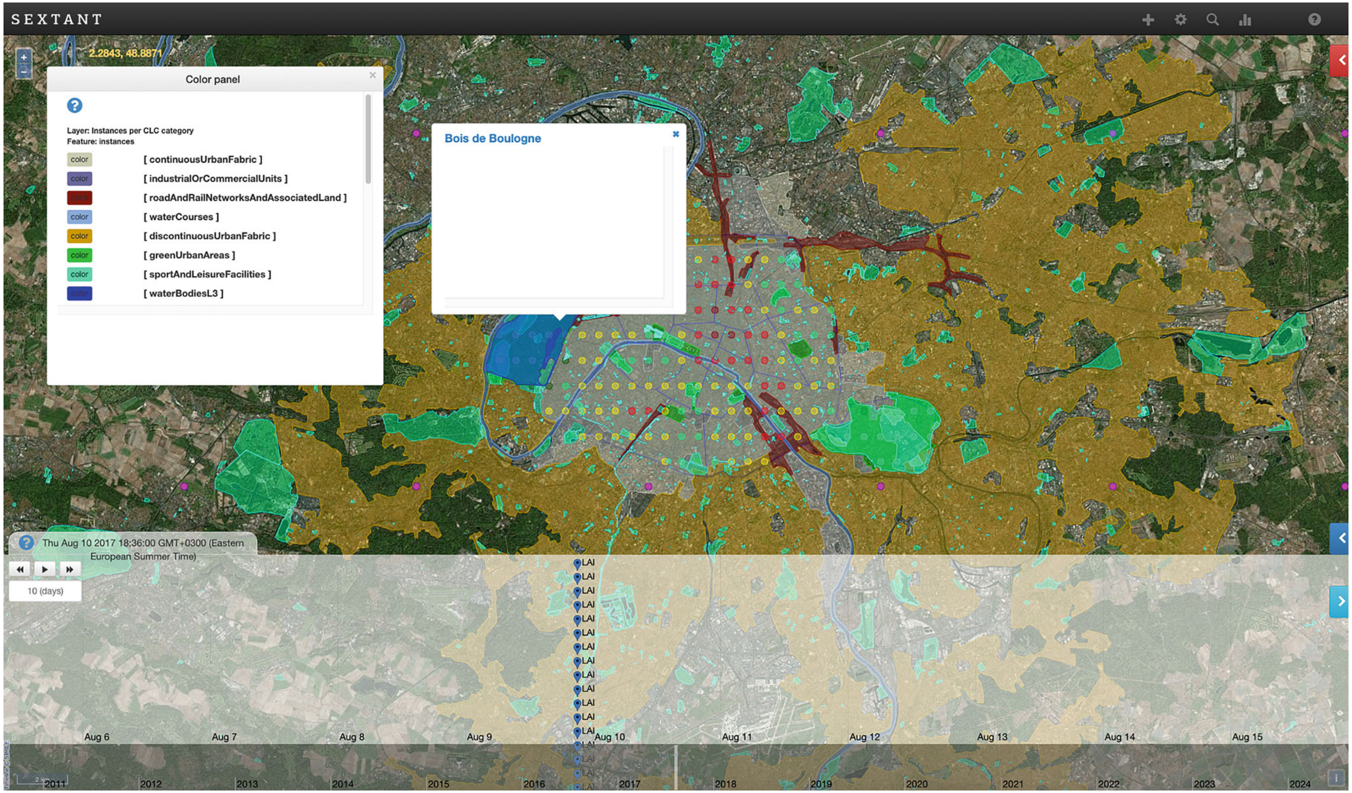
4.8 Exploration and Visualization

In order to visualize the Green City use case, we used the tool Sextant to create a map for the city of Paris, France. We used Sextant to build a temporal map that shows the "greenness" of Paris, using the datasets LAI, GADM, CLC 2018, NO_2 and OSM. We show how the LAI values change over time in each administrative area of Paris and correlate these readings with the land cover of each area taken from the CORINE land cover dataset. This allows us to explain the differences in LAI values over different areas. For example, Paris areas belonging to the CORINE land cover class *clc:greenUrbanAreas* overlap with parks in OpenStreetMap and show higher LAI values over time than industrial areas.

Sextant allows us to pose GeoSPARQL/stSPARQL queries to SPARQL endpoints and visualize the results as layers on the map. Utilizing this feature, we created the thematic map for Paris[26] shown in Fig. 4, which consists of six layers:

[25] http://silk.wbsg.de.

[26] http://test.strabon.di.uoa.gr/SextantOL3/?mapid=mpm2tf7ha6ai5f78_.


Color panel

Layer: Instances per CLC category
Feature: instances

color	[continuousUrbanFabric]
color	[industrialOrCommercialUnits]
color	[roadAndRailNetworksAndAssociatedLand]
color	[waterCourses]
color	[discontinuousUrbanFabric]
color	[greenUrbanAreas]
color	[sportAndLeisureFacilities]
color	[waterBodiesL3]

Bois de Boulogne

Thu Aug 10 2017 18:36:00 GMT+0300 (Eastern European Summer Time)

10 (days)

Fig. 4 A temporal map illustrating the “greenness” of Paris, created in Sextant

- *GADM Paris.* This layer shows us the different divisions of Paris and how "green" each part of the city is.
- *Instances per CLC category.* This layer shows us the different CLC classes that spatially intersect with the divisions on the city.
- *LAI.* This temporal layer consists of the different mean LAI values for area of Paris, for the months June–August 2017. The dots on the map are the centroids of 300 × 300 m areas that correspond to the pixel of the satellite image that contains the observation.
- *Mean LAI per Administrative Unit.* This is a statistical visualization layer, that shows us the mean LAI value for the time period of observation, for each division of Paris.
- *OSM Parks.* This layer shows us the parks that spatially intersect with the divisions of Paris.
- *NO_2.* This layer consists of the NO_2 mean concentration values for the area of Paris, for the observed time period.

5 Summary

We presented a data science pipeline for big, linked and open EO data and showed how this pipeline can be used to develop a Green City use case. The pipeline is implemented using the software developed in five FP7 and Horizon 2020 projects (TELEIOS, LEO, Melodies, Optique and Copernicus App Lab). The work presented in this chapter is now continued in the Horizon 2020 project ExtremeEarth, where we develop deep learning and big data techniques for Copernicus data in the context of two use cases: Food Security and Polar.

Acknowledgments This work has been funded by the FP7 projects TELEIOS (257662), LEO (611141), MELODIES (603525) and the H2020 project Copernicus App Lab (730124).

References

1. Auer, S., Bühmann, L., Dirschl, C., et al. (2012). Managing the life-cycle of linked data with the LOD2 stack. In *ISWC* .
2. Auer, S., Scerri, S., Versteden, A., Pauwels, E., Charalambidis, A., Konstantopoulos, S., Lehmann, J., Jabeen, H., Ermilov, I., Sejdiu, G., Ikonomopoulos, A., Andronopoulos, S., Vlachogiannis, M., Pappas, C., Davettas, A., Klampanos, I.A., Grigoropoulos, E., Karkaletsis, V., de Boer, V., Siebes, R., Mami, M.N., …Vidal, M. (2017). The bigdataeurope platform – supporting the variety dimension of big data. In *Web Engineering – 17th International Conference, ICWE 2017, Rome, Italy, June 5–8, 2017, Proceedings* (pp. 41–59).
3. Bereta, K., Caumont, H., Daniels, U., Goor, E., Koubarakis, M., Pantazi, D., Stamoulis, G., Ubels, S., Venus, V., & Wahyudi, F. (2019). The copernicus app lab project: Easy access to copernicus data. In *Advances in Database Technology – 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26–29, 2019* (pp. 501–511).

4. Bereta, K., & Koubarakis, M. (2016). Ontop of geospatial databases. In *The Semantic Web – ISWC 2016 – 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I* (pp. 37–52).
5. Bereta, K., Smeros, P., & Koubarakis, M. (2013). Representation and querying of valid time of triples in linked geospatial data. In *The Semantic Web: Semantics and Big Data, Lecture Notes in Computer Science* (Vol. 7882, pp. 259–274). Springer.
6. Bereta, K., Xiao, G., & Koubarakis, M. (2019). Ontop-spatial: Ontop of geospatial databases. *Journal of Web Semantics*, *58*, 100514.
7. Blower, J., Clifford, D., Goncalves, P., & Koubarakis, M.: The melodies project: Integrating diverse data using linked data and cloud computing. In *Proceedings of the 2014 Conference on Big Data from Space (BiDS)* (2014)
8. Burgstaller, S., Angermair, W., Migdall, S., Bach, H., Vlachopoulos, I., Savva, D., Smeros, P., Stamoulis, G., Bereta, K., & Koubarakis, M. (2017). Leopatra: A mobile application for smart fertilization based on linked data. In *Proceedings of the 8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2017), Chania, Crete Island, Greece, September 21–24, 2017* (pp. 160–171). http://ceur-ws.org/Vol-2030/HAICTA_2017_paper17.pdf
9. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., & Xiao, G. (2017). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, *8*(3), 471–487.
10. Das, S., Sundara, S., & Cyganiak, R. (2012). R2RML: RDB to RDF mapping language. http://www.w3.org/TR/r2rml/
11. Dimou, A., Vander, S., et al. (2014). RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web*. http://events.linkeddata.org/ldow2014/papers/ldow2014_paper_01.pdf
12. Espinoza-Molina, D., & Datcu, M. (2013). Earth-observation image retrieval based on content, semantics, and metadata. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(11), 5145–5159.
13. Espinoza-Molina, D., Nikolaou, C., Dumitru, C.O., Bereta, K., Koubarakis, M., Schwarz, G., & Datcu, M. (2015). Very-high-resolution SAR images and linked open data analytics based on ontologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(4), 1696–1708.
14. European Commission, European Space Agency. (2019). Copenicus sentinel data access annual report 2019. Available from https://earth.esa.int/web/sentinel/news/-/article/copernicus-sentinel-data-access-annual-report-2019
15. GAEL, NOA, GRNET, Serco. (2019). Copernicus market report. Available from https://www.copernicus.eu/sites/default/files/2019-02/PwC_Copernicus_Market_Report_2019_PDF_version.pdf
16. Garbis, G., Kyzirakos, K., & Koubarakis, M. (2013). Geographica: A benchmark for geospatial rdf stores (long version). In *The Semantic Web – ISWC 2013, Lecture Notes in Computer Science* (Vol. 8219, pp. 343–359). Springer.
17. Ioannidis, T., Garbis, G., Kyzirakos, K., Bereta, K., & Koubarakis, M. (2019). Evaluating geospatial RDF stores using the benchmark geographica 2. CoRR abs/1906.01933
18. Koubarakis, M., Bereta, K., Bilidas, D., Giannousis, K., Ioannidis, T., Pantazi, D., Stamoulis, G., Dowling, J., Haridi, S., Vlassov, V., Bruzzone, L., Paris, C., Eltoft, T., Krämer, T., Charalambidis, A., Karkaletsis, V., Konstantopoulos, S., Kakantousis, T., Datcu, M., Dumitru, C.O., Appel, F., . . . Fleming, A. (2019). From copernicus big data to extreme earth analytics. In *Advances in Database Technology – 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26–29, 2019* (pp. 690–693).
19. Koubarakis, M., Bereta, K., Papadakis, G., Savva, D., Stamoulis, G. (2017). Big, linked geospatial data and its applications in earth observation. *IEEE Internet Computing*, *July/August*, 87–91.
20. Koubarakis, M., Kontoes, C., & Manegold, S. (2013). Real-time wildfire monitoring using scientific database and linked data technologies. In *Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18–22, 2013* (pp. 649–660).

21. Koubarakis, M., Kyzirakos, K., Nikolaou, C., Garbis, G., Bereta, K., Dogani, R., Giannakopoulou, S., Smeros, P., Savva, D., Stamoulis, G., Vlachopoulos, G., Manegold, S., Kontoes, C., Herekakis, T., Papoutsis, I., . . . Michail, D. (2016). Managing big, linked, and open earth-observation data: Using the TELEIOS/LEO software stack. *IEEE Geoscience and Remote Sensing Magazine*, *4*(3), 23–37.
22. Koubarakis, M., Sioutis, M., Kyzirakos, K., Karpathiotakis, M., et al. (2012). Building virtual earth observatories using ontologies, linked geospatial data and knowledge discovery algorithms. In *ODBASE*.
23. Kyzirakos, K., Koubarakis, M., & Kaoudi, Z. (2009). Data models and languages for registries in SemsorGrid4Env. Deliverable D3.1, Dept. of Informatics and Telecommunications, University of Athens.
24. Kyzirakos, K., Karpathiotakis, M., & Koubarakis, M. (2012). Strabon: A semantic geospatial DBMS. In *The Semantic Web – ISWC 2012 – 11th International Semantic Web Conference, Boston, MA, USA, November 11–15, 2012, Proceedings, Part I* (pp. 295–311)
25. Kyzirakos, K., Karpathiotakis, M., & Koubarakis, M. (2012). Strabon: A Semantic Geospatial DBMS. In: *ISWC*.
26. Kyzirakos, K., Savva, D., Vlachopoulos, I., Vasileiou, A., Karalis, N., Koubarakis, M., & Manegold, S. (2018). Geotriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings. *Journal of Web Semantics*, *52–53*, 16–32.
27. Maali, F., Cyganiak, R., & Peristeras, V. (2012). A publishing pipeline for linked government data. In *ESWC*.
28. Nikolaou, C., Dogani, K., Bereta, K., Garbis, G., Karpathiotakis, M., Kyzirakos, K., & Koubarakis, M. (2015). Sextant: Visualizing time-evolving linked geospatial data. *Journal of Web Semantics*, *35*, 35–52.
29. Paris, C., Weikmann, G., & Bruzzone, L. (2020). Monitoring of agricultural areas by using Sentinel 2 image time series and deep learning techniques. In L. Bruzzone, F. Bovolo, & E. Santi (Eds.) *Image and Signal Processing for Remote Sensing XXVI* (Vol. 11533, pp. 122–131). International Society for Optics and Photonics, SPIE.
30. Perry, M., & Herring, J. (2012). Geosparql – a geographic query language for RDF data. Available from https://www.ogc.org/standards/geosparql
31. Saveta, T., Fundulaki, I., Flouris, G., & Ngomo, A. N. (2018). Spgen: A benchmark generator for spatial link discovery tools. In *The Semantic Web – ISWC 2018 – 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I* (pp. 408–423).
32. Sherif, M. A., Dreßler, K., Smeros, P., & Ngomo, A. N. (2017). Radon – rapid discovery of topological relations. In *AAAI* (pp. 175–181).