Editor: Carole Goble • *cag@cs.man.ac.uk*

Big, Linked Geospatial Data and Its Applications in Earth Observation

Manolis Koubarakis, Konstantina Bereta, George Papadakis, Dimitrianos Savva, and George Stamoulis • *National and Kapodistrian University of Athens*

If the terabytes of Earth observation data currently stored in archives are published on the web using the linked data paradigm, data discovery, integration with other data sources, and the development of applications will become much easier.

Terabytes of geospatial data have been made freely available recently on the web. For example, data are available from gazetteers such as GeoNames, maps from geospatial search engines like Google Maps and OpenStreetMap, and user-contributed content from social networks such as Foursquare.

Some particularly important rich sources of open and free geospatial data are the satellite programs of various countries, such as the Landsat program of the US and the Copernicus program of the European Union (EU). Satellite images can be used 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 because satellite data are hidden in various archives operated by NASA, the European Space Agency, and other national space agencies. Therefore, application developers need to search in these archives to discover the needed data and integrate them into their applications. In this article, we show how to break these barriers by publishing this data in the Resource Description Framework (RDF), interlinking it with other relevant data, and making it freely available on the web to enable easy development of geospatial applications.

Big, Linked, and Open EO Data Lifecycle

The life of Earth observation (EO) data starts with the data's generation in the ground segment of a satellite mission, where the management of this so-called payload data is an important activity. Figure 1 gives a high-level view of the lifecycle of big, linked EO data as we envisioned it in our work. Each phase of the lifecycle and its associated software tools is discussed in more detail in the following.

Ingestion, Processing, Cataloguing, and Archiving

Raw data, often from multiple satellite missions, are first ingested, processed, catalogued, and archived. This phase involves processing results in the creation of various standard products (Levels 1, 2, and so on, in EO jargon; raw data are Level 0), together with extensive metadata describing them.

Satellite Image Descriptor Extraction, KDD, and Semantic Annotation

We extended traditional image-processing methods to deal with the specificities of satellite images and extract image descriptors — for example, texture features or spectral characteristics of an image. Knowledge discovery and data mining (KDD) techniques combine image

1089-7801/17/$33.00 © 2017 IEEE

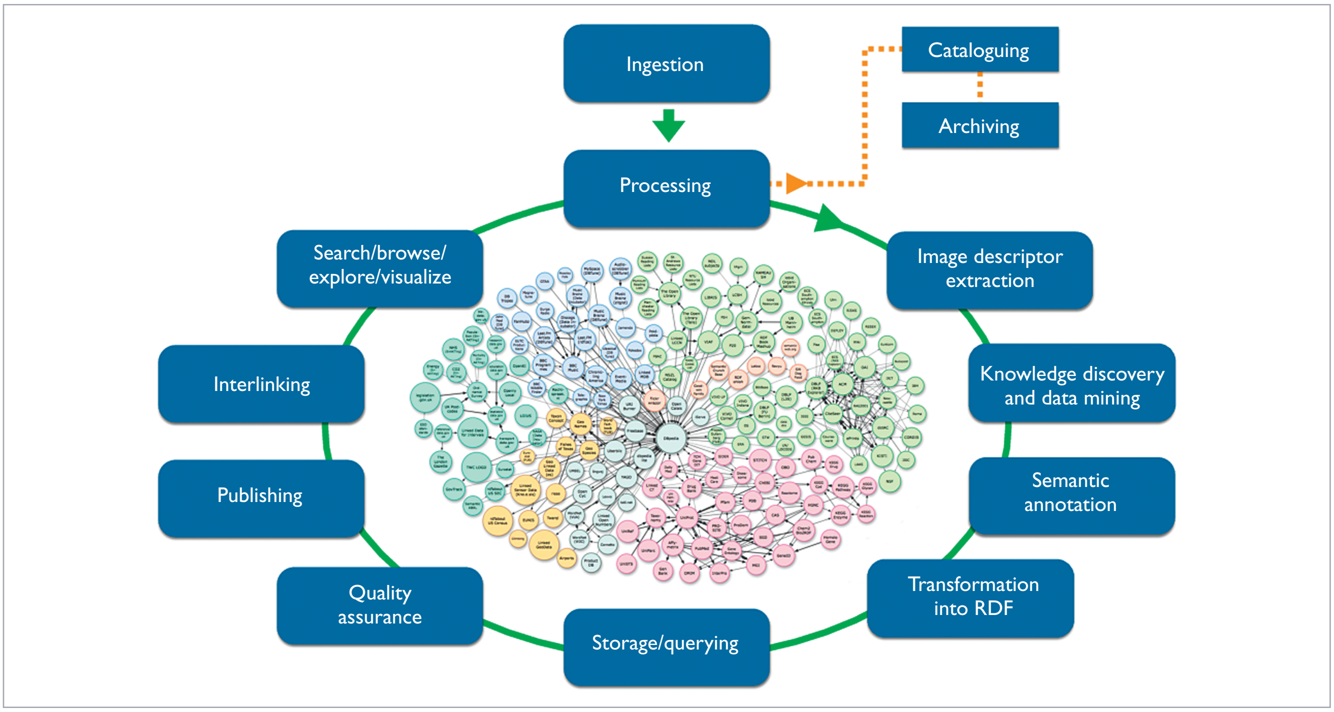

Figure 1. The lifecycle of big, linked, open Earth observation (EO) data. The yellow dashed line indicates traditional processing chains in EO datacenters. The green circle captures our additions.

descriptors, image metadata, and auxiliary data (such as GIS data) to determine concepts from a domain ontology (for example, a forest, lake, fire, or burned area) that characterize an image's content. Hierarchies of domain concepts are formalized using ontologies encoded in the Web Ontology Language, OWL-2, and are used to annotate standard products. Annotations are expressed in RDF and are made available as linked data so that they can be combined easily with other publicly available linked data sources (such as GeoNames, OpenStreetMap, and DBpedia) to allow for the expression of rich user queries.

Semantic Annotation

For encoding semantic annotations and publishing geospatial and temporal linked data, we developed the data model stRDF and the query language stSPARQL. The model stRDF is an extension to the W3C standard RDF that allows the representation of geospatial data that changes over time. It is accompanied by stSPARQL, an extension of the query language SPARQL 1.1 for querying and updating stRDF data. Both stRDF and stSPARQL use the Open Geospatial Consortium (OGC) standards Well-Known Text (WKT) and Geography Markup Language (GML) for the representation of temporal and geospatial data. Both stRDF and stSPARQL have been implemented in the system Strabon (see http://strabon.di.uoa.gr), which extends the well-known RDF store Sesame and uses PostgreSQL or MonetDB as the backend spatially- and temporally-enabled database management system. As shown by our experiments, Strabon is currently the most functional and efficiently performing geospatial and temporal RDF store available.

In our work, we use stRDF to represent satellite image metadata (for example, time of acquisition or geographical coverage), knowledge extracted from satellite images (for example, a certain image pixel is a fire hotspot), and auxiliary geospatial datasets encoded as linked data. We 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 25 August 2007 that covers the area of Peloponnese and contains hotspots corresponding to forest fires located within 2 km of a major archaeological site. Encoding this information request today in a typical interface of an EO data archive such as the Copernicus Open Access Hub (see https://scihub.copernicus.eu) is impossible, because domain-specific concepts such as "forest fires" aren't included in the archive metadata, and thus they can't be used as search criteria. In Copernicus Open Access Hub and other similar web interfaces, search criteria include a hierarchical organization of avail able products (for example, high-resolution optical or synthetic aperture radar data), together with a temporal and geographic selection menu.

Semantic Catalogue for the TerraSAR-X Archive

The workflow for constructing a semantic catalogue for the TerraSAR-X archive of Germany's Aerospace Center (DLR) can be summarized as follows. First, the TerraSAR-X products are obtained from the archive and stored separately into an image and metadata database. Then, each image is tiled into patches based on the resolution and pixel spacing extracted from the metadata database. For each patch, its quick-look is generated and stored into a quick-look database. Then, the primitive features from each tiled patch are extracted and stored into a primitive feature database. The three databases are implemented using MonetDB. Then, the features are grouped into categories from a predefined hierarchy (the DLR ontology) using an interactive learning algorithm. These categories are used to populate the semantic catalogue.

As a proof of concept, this workflow has been applied to a big dataset containing 300 scenes from the DLR TerraSAR-X archive (around 3 Tbytes of data). Applying the knowledge discovery and data mining (KDD) framework to this dataset resulted in detecting 850 semantic classes with high precision and recall.[1] As an example, Figure A shows the visualization of the information regarding the area of Venice that was introduced in the semantic catalogue using Sextant.

Reference

1. D. Espinoza-Molina et al., "Very-High-Resolution SAR Images and Linked Open Data Analytics Based on Ontologies," *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 4, 2014, pp. 1696–1708.

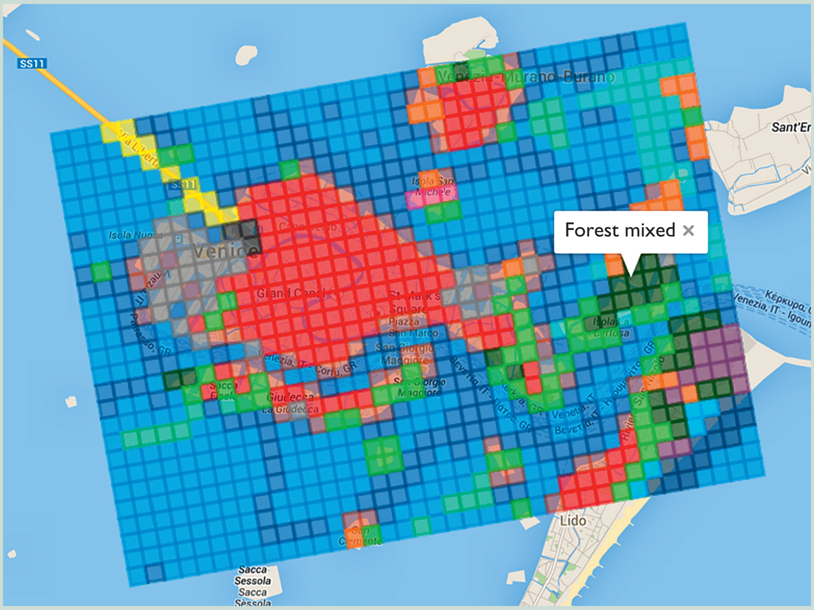

Figure A. Visualization of the information regarding the area of Venice that was introduced in the semantic catalogue using Sextant.

With the KDD techniques, we can characterize satellite image regions with concepts from appropriate ontologies (for example, land-cover ontologies with concepts such as water-body, lake, and forest; or environmental monitoring ontologies with concepts such as forest fire and flood). These concepts are encoded in OWL ontologies and are used to annotate EO products. In this way, 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, we would need to join it with information obtained from auxiliary data sources to answer the previous query. Although such open sources of data are available to EO datacenters, they aren't currently used to support sophisticated ways of end-user querying in web interfaces such as the Copernicus Open Access Hub. In our work, we assume that auxiliary data sources, especially geospatial ones, are encoded in stRDF and are available as linked geospatial data; thus, stSPARQL easily can be used to express information requests as in the previous example.

Transformation into RDF

This phase transforms vector or raster EO data from their standard formats (for example, ESRI shapefile or NetCDF) into RDF. We advanced the state of the art in transforming EO data and geospatial data into RDF by developing GeoTriples (see https://github.com/LinkedEOData/GeoTriples). GeoTriples is a tool that transforms vector data and their metadata into RDF and that natively supports many popular geospatial data formats, including shapefiles, spatially enabled DBMS, Keyhole Markup Language (KML), and GeoJSON.

Storage/Querying

This phase deals with storing all relevant EO data and metadata on persistent storage so that they're readily available for querying in subsequent phases. In our work, we use MonetDB (see www.monetdb.org) for the storage of raw image data and metadata; while we use the spatiotemporal RDF store system Strabon and the query language stSPARQL for storing/querying semantic annotations and other types of linked, geospatial data possibly originating from transforming EO products into RDF.

Often, relevant geospatial data are stored in geospatial relational databases (for example, PostGIS) and aren't available as linked data. When these databases are frequently updated and/or are large, domain experts are discouraged from transforming the data into RDF and then storing it in a triple store such as Strabon. For this reason, we developed the system Ontop-Spatial, which is a geospatial extension of the Ontology-Based Data Access system Ontop (see https://github.com/ConstantB/ontop-spatial). 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. Our experimental evaluation showed that this approach is not only simpler for users (because it doesn't require materialization of data), but is also more efficient in terms of query response time.

Quality Assurance

Before linked EO data are ready for publication, this step cleans the data by, for example, removing duplicates and so on. An important issue in this phase is entity resolution, which can also be viewed as part of the linking phase.

Publishing

This phase makes linked EO data publicly available in the linked open data (LOD) cloud using well-known data repository technologies such as the Comprehensive Knowledge Archive Network (CKAN). In this way, others can discover and share this data, avoiding duplication of effort.

Interlinking

This is an important phase in the linked EO data lifecycle, because much of linked data's value comes through connecting seemingly disparate data sources to each other. Until now, there hasn't been much research or tools for interlinking linked EO data. If we consider other published linked datasets that aren't from the EO domain, but have similar temporal and geospatial characteristics, the situation is the same. These datasets are typically linked only with `owl:sameAs` links and only to core datasets such as DBpedia or GeoNames. In addition, links are often created manually.

With our work, we advance the area of interlinking of linked open data by concentrating on the geospatial, temporal, and measurement characteristics of EO data. Specifically, we address the problem of discovering other types of geospatial or temporal semantic links. In linked EO datasets, it's often useful to discover links involving topological relationships, for example, `A geo:sfContains F` where `A` is the area covered by a remotely sensed multispectral image `I`, `F` is a geographical feature of interest (field, lake, city, and so on), 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`.

We dealt with these issues by extending the well-known link discovery tool Silk to discover precise geospatial and temporal links among spatiotemporal RDF data (see http://silkframework.org).

Search/Browse/Explore/Visualize

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

For this phase of the lifecycle, we developed the tool Sextant (see http: //sextant.di.uoa.gr). Sextant is a Web-GIS tool that produces maps by combining geospatial data from SPARQL endpoints and well-known GIS file formats. To achieve interoperability with other well-known GIS tools, Sextant is based on OGC standards for vector and raster data such as WKT, GML, KML, and GeoJSON. Sextant supports the creation of layers using the OpenGIS Web Map Service Interface Standard that's a standard protocol for serving georeferenced map images over the web, and the OGC Web Feature Service 2.0 Interface Standard that defines interfaces for describing data manipulation operations of geographic features.

Application Examples

The sidebar "A Semantic Catalogue for the TerraSAR-X Archive" showcases the lifecycle of big linked open EO data in a working application. In a related article, the lifecycle is presented with more details, and is applied to the case of wildfire monitoring using satellite images and related GIS data.[1]

More recently, the tools presented here were used in the Big Data Europe project (see www.big-data-europe.eu) to develop a pilot application in the area of space and security, The pilot aims to enhance the process of detecting changes in land cover or land use from satellite images (for example, the construction or destruction of settlements) and correlating them with the detection of geolocated events in news sites and social media. Interweaving remote sensing with social sensing constitutes a key advancement in the space and security domain, where useful information can be derived not only from EO products, but also from their combination with news articles and the user-generated content from social media. In the pilot, the tool GeoTriples is used to transform geospatial data into RDF, Strabon is used to store linked geospatial data, and Sextant has been extended to function as a user-friendly graphical interface for the whole application.

Big, linked, and open EO data can be managed using the technologies developed in the TELEIOS and Linked Open Earth Observation Data (LEO) projects. Our group's work presented here concentrates more on linked open data and has only partially addressed big data issues. The area of big data is where our current work concentrates. We're reengineering GeoTriples, Strabon, and Ontop-Spatial to take advantage of big data technologies Apache Hadoop and Spark and their recent extensions for big, geospatial data. All the tools presented here (Strabon, Ontop-Spatial, GeoTriples, Silk and Sextant) are available as open source.

Acknowledgments

This work has been funded by European Union Seventh Framework Programme (FP7) projects TELEIOS (257662), LEO (611141), Melodies (603525), and Horizon 2020 projects BigDataEurope (730124) and Copernicus App Lab (730124).

Reference

1. M. Koubarakis et al., "Managing Big, Linked, and Open Earth-Observation Data: Using the Teleios/LEO Software Stack," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 3, 2016 pp. 23–37.

Manolis Koubarakis is a professor at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens. His research interests include artificial intelligence, the Semantic Web, and linked data. Koubarakis has a PhD in computer science from the National Technical University of Athens. Contact him at koubarak@di.uoa.gr.

Konstantina Bereta is a PhD student at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens. Her research interests include the Semantic Web and linked data. Bereta has an MSc in informatics and telecommunications from the National and Kapodistrian University of Athens. Contact her at konstantina.bereta@di.uoa.gr.

George Papadakis is a postdoctoral fellow at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens. His research interests include artificial intelligence, databases, and the web. Papadakis has a PhD in computer science from the Leibniz University of Hanover. Contact him at gpapadis@di.uoa.gr.

Dimitrianos Savva is a research assistant at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens. His research interests include the Semantic Web and linked data. Savva has an MSc in informatics and telecommunications from the National and Kapodistrian University of Athens. Contact him at dimis@di.uoa.gr.

George Stamoulis is a PhD student at the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens. His research interests include the Semantic Web and linked data. Stamoulis has an MSc in informatics and telecommunications from the National and Kapodistrian University of Athens. Contact him at gstam@di.uoa.gr.