Linked Earth Observation Data: The Projects TELEIOS and LEO

MANOLIS KOUBARAKIS†, KOSTIS KYZIRAKOS†, CHARALAMPOS NIKOLAOU†, GEORGE GARBIS†, KONSTANTINA BERETA†, PANAYIOTIS SMEROS†, STELLA GIANAKOPOULOU†, KALLIROI DOGANI‡, MARIA KARPATHIOTAKI‡, IOANNIS VLACHOPOULOS‡
†National and Kapodistrian University of Athens, Greece
‡Centrum Wiskunde & Informatica, Amsterdam, NLD
{koubarak,charnik,ggarbis, Konstantina.Bereta, psmeros, sgian, kalliroi, mkparat, johnvl}@di.uoa.gr
Kostis.Kyzirakos@cwi.nl

ABSTRACT
Lots of Earth Observation (EO) data has become available at no charge in Europe and the US recently and there is a strong push for more open EO data. Open EO data that are currently made available by space agencies are not following the linked data paradigm. Therefore, from the perspective of a user, the EO data and other kinds of geospatial data necessary to satisfy his or her information need can only be found in different data silos, where each silo may contain only part of the needed data. Opening up these silos by publishing their contents as RDF and interlinking them with semantic connections will allow the development of data analytics applications with great environmental and financial value.

In this paper, we present the advances in the areas of Semantic Web and Linked Data achieved by the European project TELEIOS and give a short overview of the goal of the new European project LEO.

1. INTRODUCTION

Lots of Earth Observation (EO) data has become available at no charge in Europe and the US recently and there is a strong push for more open EO data. For example, a recent paper on Landsat data use and charges by the US National Geospatial Advisory Committee - Landsat Advisory Group starts with the following overarching recommendation: “Landsat data must continue to be distributed at no cost”.

Similarly, the five ESA Sentinel satellites that would soon go into orbit have already adopted a fully open and linked data access policy.

Linked data is a new data paradigm which studies how one can make RDF data available on the Web, and interconnect it with other data with the aim of increasing its value [4].

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 [8]. 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). The abundance of this data can prove useful to the new missions (e.g., 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, open EO data that are currently made available by space agencies such as ESA and NASA are not following the linked data paradigm. Therefore, from the perspective of a user, the EO data and other kinds of geospatial data necessary to satisfy his or her information need can only be found in different data silos, where each silo may contain only part of the needed data. Opening up these silos by publishing their contents as RDF and interlinking them with semantic connections will allow the development of data analytics applications with great environmental and financial value.

The European project TELEIOS1 is 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. Examples of such applications include wildfire monitoring and burnt scar mapping, semantic catalogues for EO archives, and rapid mapping. The wildfire monitoring application is available on the Web2 and has been used operationally by government agencies in Greece in the summer fires of 2012 and 2013. Recently, it has also been awarded 3rd place in the Semantic Web Challenge.

TELEIOS concentrated on developing data models, query languages, scalable query evaluation techniques, and efficient data management systems that can be used to prototype applications of linked EO data. However, developing a methodology and related software tools that support the whole lifecycle of linked open EO data (e.g., publishing, interlinking etc.) has not been tackled by TELEIOS. The main objective of the new European project “Linked Open Earth Observation Data for Precision Farming” (LEO) presented in this paper is to go beyond TELEIOS by designing and implementing software supporting the whole life cycle of linked open EO data and its combination with linked geospatial data, and by developing a precision farming application that heavily utilizes such data.

The rest of the paper is organized as follows. Section 2 presents in detail some of the research that has been carried in TELEIOS and will be carried out in LEO. The discussion on the whole life cycle of linked open EO data, serves as a short introduction to the whole research agenda of LEO while presenting the TELEIOS achievements in the respec-

1http://www.earthobservatory.eu/
2http://papos.space.noa.gr/fend_static/
2. THE LIFE CYCLE OF LINKED OPEN EO DATA

Developing a methodology and related software tools that support the whole life cycle of linked open EO data has not been tackled by any research project in the past, although there is plenty of such work for linked data e.g., by project LOD2 and others [1, 10]. Capturing the life cycle of open EO data and the associated entities, roles and processes of public bodies making available this data is the first step in achieving LEO’s main objective of bringing the linked data paradigm to EO data centers, and re-engineering the life cycle of open EO data based on this paradigm.

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 1 gives a high-level view of the life cycle of linked EO data as we envision it at this moment in project LEO (this is a preliminary version which will be further refined in the course of the project).

Let us now briefly discuss each one of these phases.

2.1 Ingestion, processing, cataloguing and archiving

Raw data, often from multiple satellite missions, is ingested, processed, cataloged 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. For example, in the fire monitoring application developed in project TELEIOS, images from the SEVIRI sensor are processed (cropped, georeferenced and run through a pixel classification algorithm) to detect pixels that are hotspots. Then these pixels are stored as standard products in the form of shapefiles. Raw data and derived products are complemented by auxiliary data, e.g., various kinds of geospatial data such as maps, land use/land cover data, etc.

Raw data, derived products, metadata and auxiliary data are stored in various storage systems and are made available using a variety of policies depending on their volume and expected future use. For example, in the TerraSAR-X archive of DLR, long term archiving is done using a hierarchy of storage systems (including a robotic tape library), which offers batch to near-line access, while product metadata are available on-line by utilizing a relational DBMS and an object-based query language.

2.2 Transformation into RDF

This phase transforms vector or raster EO data from their standard formats (e.g., shapefiles or GeoTIFF) into RDF. TELEIOS has developed two technologies that are important for the first two of the phases (ingestion and processing): the SciQL data model and query language [7] and data vaults [6]. SciQL is an SQL-based query language for scientific applications with arrays as first class citizens [7]. It allows stating complex satellite image processing functions as declarative SciQL queries, thus it eases substantially the development of processing chains run by EO data centers today. The data vault is a mechanism that provides a true symbiosis between a DBMS and existing (remote) file-based repositories such as the ones used in EO applications [6]. The data vault keeps the data in its original format and place, while at the same time enables transparent data and metadata access and analysis using the SciQL query language. SciQL and the data vault mechanism are implemented in the well-known column store MonetDB3.

2.3 Transformation into RDF

In the DLR knowledge extraction, knowledge discovery and data mining framework developed in TELEIOS [11], traditional raw data processing has been augmented with content extraction methods that deal with the specificities of satellite images and derive image descriptors (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., forest, lake, fire, burned area) that characterize the content of an image. Hierarchies of domain concepts are formalized using OWL ontologies and are used to annotate standard products. Annotations are expressed in RDF 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.

For this purpose we developed an RDFS ontology that captures the contents of the TerraSAR-X datasets. The developed ontology4 comprises the following two major classification schemes: the product classification scheme, which consists of all the different product-related concepts included in the dataset (e.g., the concept of an image, a patch, a label, a feature vector, a product) and the land cover/use classification scheme for annotating image patches.

Similarly, to annotate semantically standard products produced by the hotspot detection processing chain of NOA, we have developed an appropriate ontology5. The main classes of the NOA ontology are RawData, Shapefile, and Hotspot, which represent files with raw data (e.g., sensor measurements), ESRI shapefiles, which are the outputs of the hotspot detection processing chain and hotspots, which are extracted from shapefiles, respectively. For interoperability purposes, these classes have been defined as sub-classes of corresponding classes of the SWEET6 ontology.

3http://www.monetdb.org/
4http://www.earthobservatory.eu/ontologies/dlrOntology-v2.owl
5http://www.earthobservatory.eu/ontologies/noaOntology.owl
6http://sweet.jpl.nasa.gov/ontology/
In TELEIOS we will advance the state of the art in transforming EO data and geospatial data into RDF by first developing a generic stand-alone tool that will be able to deal with vector data and their metadata, and to support natively all popular geospatial data formats (shape files, KML and GeoJSON initially). The tool will produce RDF data modeled as in the recent works on stSPARQL and GeoSPARQL where new data types to encode the geometry of features have been defined. Since the transformation of raster data (e.g., raw satellite images) into RDF does not appear to be reasonable, this stand-alone tool will allow the transformation of the accompanying metadata only in such cases. As an advanced alternative, we will also integrate the extraction and transformation functionality of the stand-alone tool into MonetDB, a DBMS that supports both RDF (via relational mapping using the Strabon front-end developed in TELEIOS) and arrays (natively via SciQL). This approach allows the use of SciQL during the mapping process, e.g., to extract features from the raw raster data that can then be transformed into and published as RDF. Also, it opens up possibilities for on-demand extraction and transformation when querying the RDF data using the data vault machinery of MonetDB.

2.4 Storage/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 TELEIOS, MonetDB (with SciQL and the data vault) is used for the storage of raw image data and metadata [6] while the spatiotemporal RDF store system Strabon and the query language stSPARQL is used for storing/querying semantic annotations and other kinds of linked geospatial data originating from transforming EO products into RDF.

In TELEIOS we developed the data model stRDF and the query language stSPARQL [9, 3] which are extensions of the standards RDF and SPARQL 1.1 respectively for representing and querying geospatial data that changes over time. stRDF and stSPARQL use the widely adopted OGC standards Well Known Text and Geography Markup Language to represent geospatial data. Spatial literals can be used as objects of a triple to represent the serialization of a geometric object. The temporal dimension of stRDF and stSPARQL assumes a discrete time line and uses the value space of the datatype xsd:dateTime of XML-Schema to model time. Two kinds of time primitives are supported: time instants and time periods. Temporal literals can be used as objects of a triple to represent user-defined time. In addition, they can be used to represent valid times of temporal triples.

The query language stSPARQL extends SPARQL 1.1 with functions that take as arguments spatial or temporal terms and can be used in the SELECT, FILTER, and HAVING clause of a SPARQL 1.1 query. We use functions from the “OpenGIS Simple Feature Access - Part 2: SQL Option” (OGC-SFA) standard for querying stRDF data. Similarly, we have defined a Boolean SPARQL extension function for each topological relation defined in OGC-SFA (topological relations for simple features), Egenhofer relations and RCC-8 relations. In this way stSPARQL supports multiple families of topological relations our users might be familiar with. Using these functions stSPARQL can express spatial selections, spatial joins and spatial aggregations. The stSPARQL extension functions can also be used in the SELECT clause of a SPARQL query. As a result, new spatial literals can be generated on the fly during query time based on pre-existing spatial literals. Update operations are also supported in stSPARQL conforming to the W3C standard SPARQL Update 1.1.

The query language stSPARQL is also enabled with valid time support as follows. First, temporal triple patterns are introduced as the most basic way of querying temporal triples. A temporal triple pattern is an expression of the form \( \text{spot} \), where \( \text{sp} \) is a triple pattern and \( \text{t} \) is a time period or a variable. Second, temporal extension functions are defined in order to express temporal relations between expressions that evaluate values of the datatypes xsd:dateTime and strdf:period. The first set of such temporal functions are 13 Boolean functions that correspond to the 13 binary relations of Allen’s Interval Algebra. stSPARQL offers nine functions that are “syntactic sugar”, i.e., they encode frequently-used disjunctions of these relations. stSPARQL also defines functions that allow relating an instant with a period and functions that construct new (closed open) periods from existing ones, as well as temporal aggregates.

A complete reference of the spatial and temporal extension functions of stSPARQL can be found online.7 stSPARQL and the recent OGC standard GeoSPARQL has been developed independently at about the same time, and have concluded with very similar representational and querying constructs. Both approaches represent geometries as literals of an appropriate datatype which may be encoded in various formats like GML, WKT etc. Both approaches map spatial predicates and functions that support spatial analysis to SPARQL extension functions. GeoSPARQL goes beyond stSPARQL in that it allows binary topological relations to be used as RDF properties anticipating their possible utilization by spatial reasoners (this is the topological extension and the related query rewrite extension of GeoSPARQL). In our group, such geospatial reasoning functionality has been studied in the more general context of “incomplete information in RDF”. Since stSPARQL has been defined as an extension of SPARQL 1.1, it differs from GeoSPARQL as follows. First, it offers geospatial aggregate functions and update statements that have not been considered at all by GeoSPARQL. Second, GeoSPARQL imposes an RDFS ontology for the representation of features and geometries. On the contrary, stRDF only asks that a specific literal datatype is used and leaves the responsibility of developing any ontology to the users. Finally, stSPARQL offers the capability to query the valid time dimension of triples as well as a wide set of temporal operations, while GeoSPARQL does not deal with time at all.

In TELEIOS we designed and developed the semantic spatiotemporal RDF store Strabon8 that is a full implementation of stRDF and stSPARQL. Strabon [9, 3] can be used to store linked geospatial data that change over time and evaluate stSPARQL and GeoSPARQL queries. Strabon supports spatial datatypes enabling the serialization of geometric objects in OGC standards WKT and GML. It also offers spatial and temporal selections, spatial and temporal joins, a rich set of spatial functions similar to those offered by geospatial relational database systems and support for multiple Coor-

7http://www.strabon.di.ua.gr/stSPARQL
8http://strabon.di.ua.gr/
We discuss in more detail in the “linking” phase below.

Our work on RDF literature, we developed a semantics for RDF language. Following ideas from the incomplete information presssed by a quantifier-free formula of a first-order constraint allows partial information regarding property values to be ex-

We also developed the benchmark Geographica[5] that can be used for the evaluation of the new generation of RDF stores supporting the query languages GeoSPARQL and stSPARQL. Geographica [5] is composed by two workloads with their associated datasets and queries: a real-world workload based on publicly available linked data sets and a synthetic workload. The real-world workload uses publicly available linked geospatial data, covering a wide range of geometry types (e.g., points, lines, polygons). For this workload we define a micro benchmark that tests primitive spatial functions and a macro benchmark that tests the performance of the selected RDF stores in typical application scenarios like reverse geocoding, map search and browsing, and a real-world use case from the Earth Observation do-

In the context of TELEIOS, we also extended RDF with the ability to define a new kind of literals for each datatype that can be used to represent values of properties that exist but are unknown or partially known. In the proposed extension of RDF, called RDF$^3$ [13, 14], such literals are allowed to appear only in the object position of triples. RDF$^3$ allows partial information regarding property values to be ex-

2.5 Data cleaning

Before linked EO data is ready for publication, this step is used to clean the data by e.g., removing duplicates etc. An important issue in this phase is entity resolution which we discuss in more detail in the “linking” phase below.

2.6 Publication

This phase makes linked EO data publicly available in the LOD cloud 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. In the course of developing the EO applications in TELEIOS, the following data sets were published:

- The CORINE Land Cover of Europe dataset. The CORINE Land Cover project$^9$ is an activity of the European Environment Agency that collects data regarding the land cover of European countries.
- The Urban Atlas of Europe dataset. Urban Atlas$^{10}$ is an activity of the European Environment Agency that pro-
- The Greek Administrative Geography dataset. This dataset describes the administrative divisions of Greece (prefecture, municipality, district, etc.). It has been populated with relevant data that are available in Greek open govern-
- The Coastline of Greece dataset. This dataset describes the geometry of the coastline of Greece.

All these datasets are publicly available and can be easily accessed, downloaded and queried through the Greek Linked Open Data portal$^{12}$.

2.7 Interlinking

This is a very important phase in the linked EO data life cycle since a lot of the value of linked comes through connecting seemingly disparate data sources to each other. Up to now, there has not been much research or tools for interlinking linked EO data. If one considers other published linked data sets that are not from the EO domain, but have similar temporal and geospatial characteristics, the situation is the same. These data sets are typically linked only with owl:sameAs links and only to core datasets such as DBpedia or Geonames. In addition, these links are often created manually since existing tools such as Google Refine and Silk have not been found to perform satisfactorily for these datasets [15].

In LEO we will advance the state of the art in the area of interlinking of linked open data by concentrating on the geospatial, temporal and measurement characteristics of EO data. The first problem to be studied will be entity resolution. For geospatial data, entity resolution has been studied only for location (point) datasets. We will extend the relevant techniques to the case of more complex geometries captured by the spatial literal data types of stSPARQL and GeoSPARQL that will be utilized in the data published by LEO as discussed above. If needed, we might also use ontology alignment techniques to deal with situations where the techniques of [8] would fail (e.g., when the types of features considered are synonyms or one type is a subclass of the other etc.). Finally, we will consider geospatial entity

9http://www.eea.europa.eu/publications/CORINE-landcover/
11http://geodata.gov.gr/
12http://linkedopendata.gr/
resolution among EO datasets published by LEO and already existing geospatial datasets that do not follow the stSPARQL/GeoSPARQL modelling paradigm and use different vocabularies such as W3C Geo (e.g., OpenStreetMap data published by the LinkedGeoData project). This will result in the development of techniques for geospatial entity resolution in datasets that use heterogeneous geospatial vocabularies.

We will also study the problem of discovering other kinds of semantic links that are geospatial or temporal in nature. For example, in linked EO datasets, it will often be important to discover links involving topological relationships e.g., 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 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.

Such information can be used for qualitative spatial reasoning which relies on qualitative abstractions of spatial aspects of the common-sense background knowledge. In TELEIOS we have developed PyRCC8 [16], an open source, efficient reasoner for RCC-8 written in pure Python. PyRCC8 was found to exhibit excellent experimental results and has served as the basis for a reasoner for chordal RCC-8 networks, called PyRCC8-TRIANGLE, that we have also developed within TELEIOS taking on a different research direction. PyRCC8-TRIANGLE further boosts the consistency checking process of sparse RCC-8 networks by making their underlying constraint graphs chordal rather than complete as is the case with the state of the art reasoners including PyRCC8.

2.8 Search/Browse/Explore/Visualize

This is also a very important phase since it enables users to find and explore the data they need, and start developing interesting applications. In TELEIOS we designed and developed the tool SexTant\footnote{http://sextant.di.uoa.gr/}. SexTant\footnote{http://deepenandlearn.esa.int/tiki-index.php?page=RARE\%20Project} [12, 2] is a web-based tool that makes the task of map creation and sharing, as well as the task of browsing linked spatiotemporal data easy. Similar to well-known GIS tools (e.g., ArcGIS, QGIS), SexTant can be used to produce thematic maps by layering spatiotemporal information which exists in a number of data sources ranging from standard SPARQL endpoints, to SPARQL endpoints following OGC standards for the modelling and querying of geospatial information (i.e., GeoSPARQL), or even other standards or file formats, such as KML and GeoJSON. The feature that distinguishes SexTant from other semantic web or GIS tools is that map creation and sharing, as well as exploration of data can be done in a declarative way using the query languages stSPARQL or GeoSPARQL. In this sense, SexTant is able to create useful thematic maps by layering information coming from the evaluation of SPARQL queries. For this phase in LEO, we plan to extend the tools developed in ESA project RARE\footnote{http://deepenandlearn.esa.int/tiki-index.php?page=RARE\%20Project} and the tool SexTant with additional functionalities. Finally, we plan to make these tools available for mobile devices (tablets, smartphones) to enable the use of EO data by ordinary users and application specialists alike.

3. CONCLUSIONS

We discussed the results of the project TELEIOS in the areas of Semantic Web and Linked data and how we plan to extend them in the new European project LEO which intends to bring the linked open data paradigm to EO.

4. REFERENCES