Building Virtual Earth Observatories Using Ontologies, Linked Geospatial Data and Knowledge Discovery Algorithms*

Manolis Koubarakis[1], Michael Sioutis[1], George Garbis[1], Manos Karpathiotakis[1], Kostis Kyzirakos[1], Charalampos Nikolaou[1], Konstantina Bereta[1], Stavros Vassos[1], Corneliu Octavian Dumitru[2], Daniela Espinoza-Molina[2], Katrin Molch[2], Gottfried Schwarz[2], and Mihai Datcu[2]

[1] National and Kapodistrian University of Athens, Greece
koubarak@di.uoa.gr
[2] German Aerospace Center (DLR), Germany

Abstract. Advances in remote sensing technologies have allowed us to send an ever-increasing number of satellites in orbit around Earth. As a result, satellite image archives have been constantly increasing in size in the last few years (now reaching petabyte sizes), and have become a valuable source of information for many science and application domains (environment, oceanography, geology, archaeology, security, etc.). TELEIOS is a recent European project that addresses the need for scalable access to petabytes of Earth Observation data and the discovery of knowledge that can be used in applications. To achieve this, TELEIOS builds on scientific databases, linked geospatial data, ontologies and techniques for discovering knowledge from satellite images and auxiliary data sets. In this paper we outline the vision of TELEIOS (now in its second year), and give details of its original contributions on knowledge discovery from satellite images and auxiliary datasets, ontologies, and linked geospatial data.

1 Introduction

Advances in remote sensing technologies have enabled public and commercial organizations to send an ever-increasing number of satellites in orbit around Earth. As a result, Earth Observation (EO) data has been constantly increasing in volume in the last few years, and it is currently reaching petabytes in many satellite archives. For example, the multi-mission data archive of the TELEIOS partner German Aerospace Center (DLR) is expected to reach 2 PB next year, while ESA estimates that it will be archiving 20 PB of data before the year 2020. As the volume of data in satellite archives has been increasing, so have the scientific and commercial applications of EO data. Nevertheless, it is estimated that up to 95% of the data present in existing archives has never been accessed, so the potential for increasing exploitation is very big.

* This work has been funded by the FP7 project TELEIOS (257662).

R. Meersman et al. (Eds.): OTM 2012, Part II, LNCS 7566, pp. 932–949, 2012.

TELEIOS[1] is a recent European project that addresses the need for scalable access to PBs of Earth Observation data and the effective discovery of knowledge hidden in them. TELEIOS started on September 2010 and it will last for 3 years. In the first one and a half years of the project, we have made significant progress in the development of state-of-the-art techniques in Scientific Databases, Semantic Web and Image Mining and have applied them to the management of EO data.

The contributions of this paper are the following:

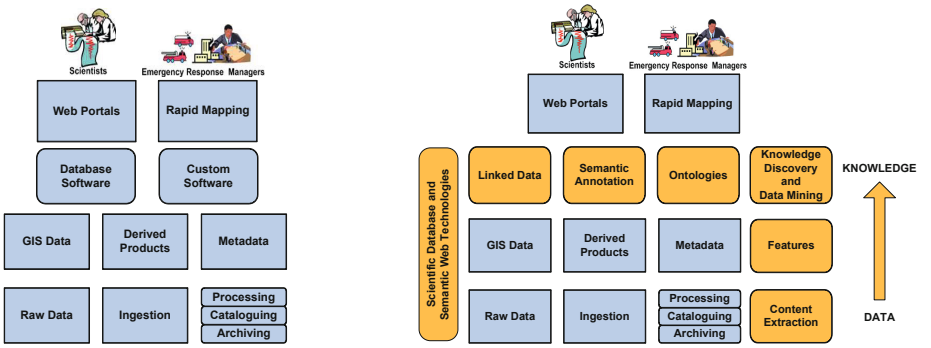
- We outline the vision of TELEIOS and explain in detail why it goes beyond operational systems currently deployed in various EO data centers. The vision of TELEIOS is also been presented in [10].
- We discuss the knowledge discovery framework developed in TELEIOS and give details of its application to radar images captured by TerraSAR-X, one of the satellites deployed by the TELEIOS partner German Aerospace Data Center (DLR). TerraSAR-X is a synthetic aperture radar (SAR) satellite launched on June 2007 in order to supply high quality radar data for the scientific observation of the Earth.
- We present briefly the data model stRDF and its query language stSPARQL which are used in TELEIOS for representing knowledge extracted from satellite images and other geospatial data sets and integrating it with other linked geospatial data sources.
- We show the added value of the TELEIOS Virtual Earth Observatory in comparison with existing EO portals such as EOWEB-NG and EO data management systems such as DIMS [28]. The added value comes from extracting knowledge from the images, encoding this knowledge in stRDF semantic annotations, and integrating with other relevant data sources available as linked data. Another interesting application of TELEIOS not discussed in this paper is the fire monitoring service presented in [14].

The rest of the paper is organized as follows. Section 2 discusses the concepts on which TELEIOS is based and explains why it improves the state-of-the-art in information systems for EO data centers. Section 3 discusses the topic of knowledge discovery from EO images. Section 4 presents the data models stRDF and stSPARQL. Section 5 presents our vision of a Virtual Earth Observatory that goes beyond existing EO portals by enabling queries that capture the semantics of the content of the images. Last, Section 6 discusses related work and Section 7 concludes the paper.

2 Basic Concepts of the TELEIOS Earth Observatory

Satellite missions continuously send to Earth huge amounts of EO data providing snapshots of the surface of the Earth or its atmosphere. The management of the so-called *payload data* is an important activity of the ground segments of

[1] `http://www.earthobservatory.eu/`



(a) Concept view of a state-of-the-art EO data center

(b) Concept view of the TELEIOS Earth Observatory

Fig. 1. Pre-TELEIOS EO data centers and the TELEIOS Virtual Earth Observatory

satellite missions. Figure 1(a) gives a high-level view of some of the basic data processing and user services available at EO data centers today, e.g., at the German Remote Sensing Data Center (DFD) of TELEIOS partner DLR through its Data Information and Management System (DIMS) [28].

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. 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 managed by DFD, 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 [28].

EO data centers such as DFD also offer a variety of user services. For example, for scientists that want to utilize EO data in their research, DFD offers the Web interface EOWEB-NG[2] for searching, inspection and ordering of products. Space agencies such as DLR and NOA might also make various other services available aimed at specific classes of users. For example, the Center for Satellite Based Crisis Information (ZKI)[3] of DLR provides a 24/7 service for the rapid provision, processing and analysis of satellite imagery during natural and environmental disasters, for humanitarian relief activities and civil security issues worldwide. Similar emergency support services for fire mapping and damage assessment are offered by NOA through its participation in the GMES SAFER program.

[2] `https://centaurus.caf.dlr.de:8443/`
[3] `http://www.zki.dlr.de/`

The TELEIOS advancements to today's state of the art in EO data processing are shown in yellow boxes in Figure 1(b) and can be summarized as follows:

- Traditional raw data processing is augmented by *content extraction* methods that deal with the specificities of satellite images and derive image descriptors (e.g., texture features, spectral characteristics of the image, etc.). 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, port, etc.) that characterize the content of an image [7].
- 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 [2] so that they can be easily combined with other publicly available linked data sources (e.g., GeoNames, LinkedGeoData, DBpedia) to allow for the expression of rich user queries.
- Web interfaces to EO data centers and specialized applications (e.g., rapid mapping) can now be improved significantly by exploiting the semantically-enriched standard products and linked data sources made available by TELEIOS. For example, an advanced EOWEB-NG-like interface to EO data archives can be developed on top of a system like Strabon[4], which is based on stRDF and stSPARQL, to enable end-users to pose very expressive queries (an example is given below). Rapid mapping applications can also take advantage of rich semantic annotations and open linked data to produce useful maps even in cases where this is difficult with current technology. Open geospatial data are especially important here. There are cases of rapid mapping where emergency response can initially be based on possibly imperfect, open data (e.g., from OpenStreetMap) until more precise, detailed data becomes available[5].

In all of the above processing stages, from raw data to application development, TELEIOS utilizes scientific database and semantic web technologies as Figure 1(b) illustrates.

In the rest of this paper we showcase the advances of TELEIOS, presenting our vision of a Virtual Earth Observatory that goes beyond existing EO portals, by enabling queries that capture the semantics of the content of the images. But first, the problem of knowledge discovery from EO images in TELEIOS is discussed.

3 Knowledge Discovery from EO Images

In this section we discuss the problem of knowledge discovery from EO images and related data sets and present the approach we follow in TELEIOS.

[4] `http://www.strabon.di.uoa.gr/`

[5] Many related ideas have recently been discussed under the topic of "Open Source Intelligence".

Knowledge Discovery from images (e.g., multimedia, satellite, medical, etc.) is now a mature subarea of Image Processing and Analysis. In the standard approach, images are first processed in order to extract visual features (either local or global) and/or segments that efficiently represent the original image. Semantic labels are then assigned to the images through classification, clustering or visual matching. These semantic labels are often represented using appropriate ontologies and are stored as annotations of the image using Semantic Web technologies [27]. The problem of selecting the set of appropriate image analysis methods and the approach to relate the results with high-level concepts is commonly referred to as "bridging the semantic gap" [25]. The recent paper [26] discusses some of the state of the art in this area and reaches the optimistic conclusion that machine understanding for multimedia images is within reach.

The state of the art in machine understanding of satellite images is significantly behind similar efforts in multimedia, but some very promising work has been carried out recently often under the aspects of international space organizations such as ESA (for example the series of Image Information Mining Conferene[6]).

TELEIOS aims to advance the state of the art in knowledge discovery from satellite images by developing an appropriate knowledge discovery framework and applying it to synthetic aperture radar images obtained by the satellite TerraSAR-X of TELEIOS partner DLR.

Satellite images are typically more difficult to handle than multimedia images, since their size often scales up to a few gigabytes, and there is also a difficulty in identifying objects and features in them. For example, mining of EO images based on content analysis is very different from mining images of faces or animals, because of the different nature of actual features (e.g., eyes, ears, stripes, or wings) that have known relationships and therefore promote the differentiation of classes [5]. Moreover, in synthetic aperture radar (SAR) images that are studied in TELEIOS, additional problems arise from the fact that these images have different acquisition properties than optical images. Essentially, SAR products may look like optical images but in reality they are mathematical products that rely on delicate radar measurements.

In [7] we presented a detailed analysis of TerraSAR-X Level 1b products[7] and identified the ones that we will use for our knowledge discovery research and the Virtual Earth Observatory implementation in TELEIOS. Each TerraSAR-X product comprises a TSX XML file which defines in detail the data types, valid entries, and allowed attributes of a product, and a TSX image. Additionally, a preview of the product in GeoTIFF[8] format is given as a quick-look image georeferenced to the WGS84 coordinate reference system, and annotated with latitude/longitude coordinates. Figure 2 shows an example of a quick-look image

[6] `http://rssportal.esa.int/tiki-index.php?page=2012_ESA-EUSC`

[7] Level 1b products are operational products offered by the TerraSAR-X payload ground segment (PGS) to commercial and scientific customers.

[8] GeoTIFF is an extension of the TIFF (Tagged Image File Format) standard which defines additional tags concerning map projection information.

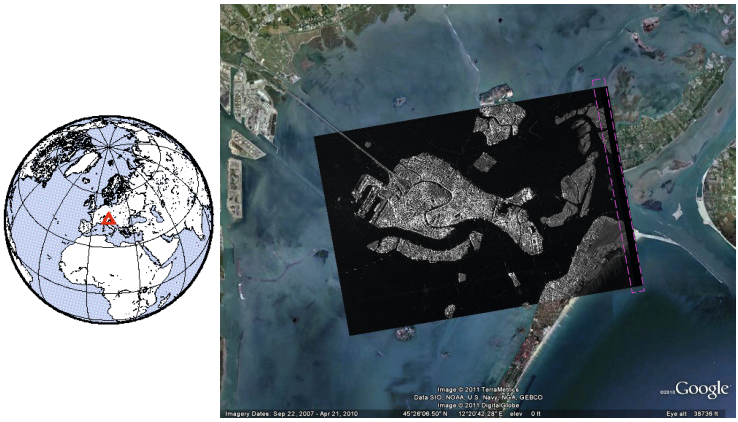


Fig. 2. Overlay on Google Earth and location of the Venice site

of Venice projected on Google Earth and its position on the globe. The quick-look image serves only as a preview of the TSX product and is not to be mistaken with the actual TSX image that is used for proccessing in the knowledge discovery framework.

The XML metadata file which is included in the delivered product packages can have sample sequences like this:

```
<productInfo>
   <missionInfo>
      <mission>TSX-1</mission>
      ...
   </missionInfo>
   <acquisitionInfo>
      ...
   </acquisitionInfo>
   ...
</productInfo>
<platform>
   <orbit>
      ...
   </orbit>
   ...
</platform>
```

We present the main steps of the knowledge discovery methodology that is currently been implemented in TELEIOS. The details of these steps are as follows:

1. *Tiling the image into patches.* In the literature of information extraction from satellite images, many methods are applied at the pixel level using a small analysis window. This approach is suitable for low resolution images but it is not appropriate for high resolution images such as SAR images from TerraSAR-X that we study in TELEIOS. Pixel-based methods cannot capture the contextual information available in images (e.g., complex structures are usually a mixture of different smaller structures) and the global features describing overall properties of images are not accurate enough. Therefore,

Table 1. Feature extraction methods

Feature extraction method	No. of features
GAFS - Gabor Filters (2 scales and 2 orientations)	48
GAFS - Gabor Filters (4 scales and 6 orientations)	8
GLCM - Gray Level Co-occurrence Matrix	48
NSFT - Nonlinear Short Time Fourier Transform	6
QMFS - Quadrature Mirror Filters (# of wavelet decompositions equal to 1)	8
QMFS - Quadrature Mirror Filters (# of wavelet decompositions equal to 2)	14

in our work, TerraSAR-X images are divided into patches and descriptors are extracted for each one. The size of the generated patches depends on the resolution of the image and its pixel spacing. Patches can be of varying size and they can be overlapping or non-overlapping. The details are discussed in [7].

2. *Patch content analysis.* This step takes as input the image patches produced by the previous step and generates feature vectors for each patch. The feature extraction methods that have been used are presented in Table 1 together with the number and kind of features they produce. The details of these methods are presented in [7].
3. *Patch classification and assignment of semantic labels.* In this step, a support vector machine (SVM) classifier is used to classify feature vectors into semantic classes. It is also possible to utilize relevance feedback from the end user to reach an improved classification. [7] presents detailed experimental results that have been obtained by applying our techniques to TerraSAR-X images to detect the 35 classes presented in Table 2. The semantic class labels are concepts from an RDFS ontology, presented in Section 5, which we have defined especially for the Virtual Earth Observatory for TerraSAR-X data.

4 The Data Model stRDF and the Query Language stSPARQL

stRDF is an extension of the W3C standard RDF that allows the representation of geospatial data that changes over time [12,15]. 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 (Well-Known Text and Geography Markup Language) for the representation of temporal and geospatial data [15].

In TELEIOS, stRDF is used to represent satellite image metadata (e.g., time of acquisition, geographical coverage), knowledge extracted from satellite images (e.g., a certain image comprises semantic annotations) and auxiliary geospatial data sets encoded as linked data. One can then use stSPARQL to express in a

Table 2. Semantic classes identified by the techniques developed by DLR

Class No.	Semantics	Class No.	Semantics
1	Bridge (type 1)	19	Forest
2	Harbor	20	Bridge (type 2)
3	River deposits	21	Water + Urban or Vegetation
4	Agriculture	22	Road + Vegetation
5	Distortions	23	Structure roof
6	Mixed Vegetation and Water	24	Train lines (type 2)
7	Vegetation	25	Urban (type 2)
8	Urban + Water	26	Grassland (rectangular shape)
9	Urban (type 1)	27	Grassland with objects
10	Cemetery	28	Building - shape
11	Water + Vegetation	29	Urban (type 3)
12	Water + Ambiguities	30	Building (reflection)
13	Water	31	Vegetation + Urban
14	Water + Boat	32	Road + Building
15	Vegetation + Building	33	Tree + Building
16	Beach area	34	Parking
17	Train line (type 1)	35	Park with street
18	Grassland		

single query an information request such as the following: "Find images containing ports near the city of Amsterdam". Encoding this information request today in a typical interface to an EO data archive such as EOWEB-NG is impossible, because information extracted from the content of the products is not included in the archived metadata, thus they cannot be used as search criteria[9]. In [3,7] we have been developing image information mining techniques that allow us to characterize satellite image regions with concepts from appropriate ontologies (e.g., landcover ontologies with concepts such as port, water-body, lake, or forest, or environmental monitoring ontologies with concepts such as forest fires, or 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, explicitly archived information.

But even if semantic information was included in the archived annotations, one would need to join it with information obtained from auxiliary data sources (e.g., maps, wikipedia etc.) to answer the above query. Although such open sources of data are available to EO data centers, they are not used currently to support sophisticated ways of end-user querying in Web interfaces such as EOWEB-NG. In TELEIOS, we assume that auxiliary data sources, especially geospatial ones, are encoded in RDF and are available as linked data, thus stSPARQL can easily

[9] In EOWEB-NG and other similar Web interfaces, search criteria include a hierarchical organization of available products (e.g., high resolution optical data, Synthetic Aperture Radar data, their subcategories, etc.) together with a temporal and geographic selection menu.

be used to express information requests such as the above. The linked data web is being populated with geospatial data quickly [1], thus we expect that languages such as stSPARQL (and the related OGC standard query language GeoSPARQL [18]) will soon be mainstream extensions of SPARQL that can be used to access such data effectively.

Let us now introduce the basic ideas of stRDF and stSPARQL. The datatypes `strdf:WKT` and `strdf:GML` are introduced for modeling geometric objects that change over time. The values of these datatypes are typed literals that encode geometric objects using the OGC standard Well-known Text (WKT) or Geographic Markup Language (GML) respectively. These literals are called *spatial literals*. The datatype `strdf:geometry` is also introduced to represent the serialization of a geometry independently of the serialization standard used. The datatype `strdf:geometry` is the union of the datatypes `strdf:WKT` and `strdf:GML`, and appropriate relationships hold for their lexical and value spaces.

Like in RDF, stRDF data is represented as triples of URIs, literals, and blank nodes in the form *"subject predicate object"*. Additionally, stRDF allows triples to have a fourth component representing the time the triple is *valid* in the domain. Since this capability has not so far been utilized in TELEIOS, we omit further discussion of this feature in this paper.

The following RDF triples encode information related to a TerraSAR-X image that is identified by the URI `dlr:Image_1.tif`. Prefix `dlr` corresponds to the namespace for the URIs that refer to the DLR applications in TELEIOS, while `xsd` and `strdf` correspond to the XML Schema namespace and the namespace for our extension of RDF, respectively.

```
dlr:Image_1.tif rdf:type dlr:Image .
dlr:Image_1.tif dlr:hasName "IMAGE_HH_SRA_spot_047.tif"^^xsd:string .
dlr:Image_1.tif strdf:hasGeometry
          "POLYGON ((12 45, 13 45, 13 46, 12 46, 12 45));
          <http://spatialreference.org/ref/epsg/4326/>"^^strdf:WKT .
dlr:Image_1.tif dlr:consistsOf dlr:Patch_1 . dlr:Patch_1 rdf:type dlr:Patch .
dlr:Patch_1 strdf:hasGeometry
        "POLYGON ((12 44, 13 44, 13 45, 12 45, 12 44))"^^strdf:WKT .
dlr:Patch_1 dlr:hasLabel dlr:Label_1 . dlr:Label_1 dlr:correspondsTo dlr:Port .
```

The third triple above shows the use of spatial literals to express the geometry of the image. This spatial literal specifies a polygon that has exactly one exterior boundary and no holes. The exterior boundary is serialized as a sequence of the coordinates of its vertexes. These coordinates are interpreted according to the WGS84 geodetic coordinate reference system identified by the URI `http://spatialreference.org/ref/epsg/4326/` which can be omitted from the spatial literal. The rest of the triples are used to annotate image `dlr:Image_1.tif` with information mined using the knowledge discovery techniques that DRL develops in the context of TELEIOS (see Section 3 for more details of these techniques). In this example, these techniques identified a port in image `dlr:Image_1.tif` (last triple above).

stSPARQL is an extension of SPARQL (the W3C standard query language for RDF data) targeted at querying geospatial information. In stSPARQL variables may refer to spatial literals (e.g., variable `?HGEO` in triple pattern `?H strdf:hasGeometry ?HGEO`). stSPARQL provides functions that can be used

in filter expressions to express qualitative or quantitative spatial relations. For example the function `strdf:contains` is used to encode the topological relation *ST_Contains* of the OGC Simple Feature Access standard[10]. stSPARQL supports also update operations (insertion, deletion, and update of stRDF triples) on stRDF data conforming to the declarative update language for SPARQL, SPARQL Update 1.1, which is a current proposal of W3C[11]. Updating stRDF data is an important requirement in TELEIOS which is utilized during post-processing of generated products for their improvement in terms of accuracy.

The following query, expressed in stSPARQL, looks for TerraSAR-X images that contain ports near the city of Amsterdam. This query can be useful for monitoring coastal zones that contain industrial ports (e.g., Rotterdam, Hamburg).

```
SELECT  ?IM ?IMNAME
WHERE { ?PR dlr:hasImage ?IM . ?IM rdf:type dlr:Image .
        ?IM dlr:hasName ?IMNAME . ?IM dlr:consistsOf ?PA .
        ?PA rdf:type dlr:Patch . ?PA strdf:hasGeometry ?PAGEO .
        ?PA dlr:hasLabel ?L . ?L rdf:type dlr:Label .
        ?L dlr:correspondsTo dlr:Port . ?A rdf:type dbpedia:PopulatedPlace .
        ?A dbpprop:name "Amsterdam"^^xsd:string . ?A geo:geometry ?AGEO .
        FILTER (strdf:distance(?PAGEO,?AGEO) < 2) . }
```

In the above query, apart from querying information for TerraSAR-X images like the one we encoded above, linked data from DBpedia[12] is also used to retrieve semantic and geospatial information about Amsterdam. DBpedia is an RDF dataset consisting of structured information from Wikipedia that allows one to link other RDF datasets to Wikipedia data. The geometries offered by DBpedia are points representing latitude, longitude and altitude information in the WGS84 coordinate reference system.

As we can see with this example, stSPARQL enables us to develop advanced semantics-based querying of EO data along with open linked data available on the web. In this way TELEIOS unlocks the full potential of these datasets, as their correlation with the abundance of data available in the web can offer significant added value.

The stRDF model and stSPARQL query language have been implemented in the system Strabon which is freely available as open source software[13]. Strabon extends the well-known open source RDF store Sesame 2.6.3 and uses PostGIS as the backend spatially-enabled DBMS. stSPARQL is essentially a subset of the recent OGC standard GeoSPARQL since it offers almost exact functionalities with the core, geometry extension and geometry topology extension of GeoSPARQL. Strabon supports this subset of GeoSPARQL as well. A detaild comparison of stSPARQL and GeoSPARQL can be found in [11,13].

Our discussion above covers only the concepts of stRDF and stSPARQL that are appropriate for understanding this paper; more details can be found in [12,15]. Let us now describe how stRDF and stSPARQL are deployed in TELEIOS by presenting their use in developing a Virtual Earth Observatory.

[10] `http://portal.opengeospatial.org/files/?artifact_id=25354`
[11] `http://www.w3.org/TR/sparql11-update/`
[12] `http://www.dbpedia.org/`
[13] `http://www.strabon.di.uoa.gr/`

5 A Virtual Earth Observatory for TerraSAR-X Data

In this section we present our initial efforts for the development of a Virtual Observatory for TerraSAR-X data and demonstrate its current functionality through a set of representative stSPARQL queries. We demonstrate that the Virtual Earth Observatory we are developing will go well beyond the current EOWEB-NG portal of DLR (which simply offers a hierarchical organization of EO products together with a temporal and geographic selection menu) to enable queries that capture the content of satellite images and its semantics. These queries exploit the full TerraSAR-X product metadata knowledge extracted from the images using the techniques of Section 3 and other auxiliary data, e.g., publicly available geospatial data, relevant GIS data, etc. These queries are of huge importance for a broad community of users because they can be used to refer to the nature and properties of the SAR products, a complicated piece of information which currently remains hidden in the archives of DLR.

We begin by introducing the reader to the ontology that describes the TerraSAR-X product metadata, the knowledge discovered from the products, and auxiliary data to be used. Then, we give some examples of stSPARQL queries that demonstrate the advantages of the Virtual Earth Observatory over the current EOWEB-NG portal deployed by DLR.

5.1 An Ontology for TerraSAR-X Data

We have developed an RDFS ontology[14] which captures the contents of the Virtual Earth Observatory. From this point on, we will refer to this ontology as the "DLR ontology".

The DLR ontology comprises the following major parts:

- The part that captures the hierarchical structure of a product and the XML metadata associated with it. Currently only a small number of metadata fields (e.g., time and area of acquisition, sensor, imaging mode, incidence angle) are included. These are the ones most often used by users of the current EOWEB portal as well as the ones utilized by our knowledge discovery techniques.
- The part that defines the RDFS classes and properties that formalize the outputs of the knowledge discovery step (e.g., patch, feature vector).
- The part that defines the land cover/use classification scheme for annotating image patches that was constructed while experimenting with the knowledge discovery framework presented in Section 3. We made the decision not to employ a "full blown" land cover/use ontology (e.g., Europe's CORINE[15]), because the annotation of image patches currently being carried out uses only simple labels. The classification scheme therefore provides a basic structure for annotating patches, a part of which is shown in Figure 3. It is clear that the top level has a general meaning (e.g., transportation, water, etc.),

[14] `http://www.earthobservatory.eu/ontologies/dlrOntology.owl`
[15] `http://harmonisa.uni-klu.ac.at/ontology/corine.owl`

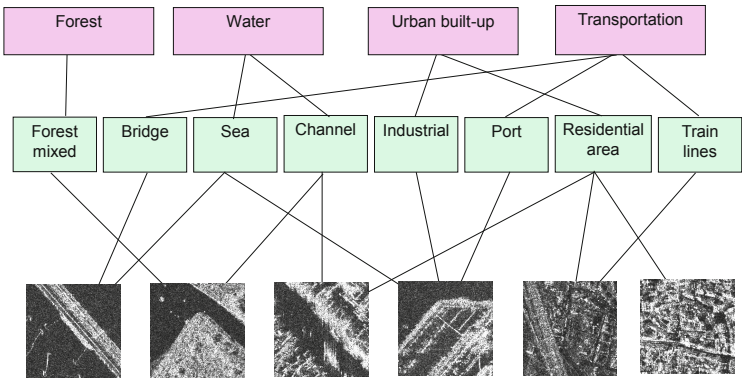


Fig. 3. Two-level concept taxonomy and example of the multi-semantic annotation of patches

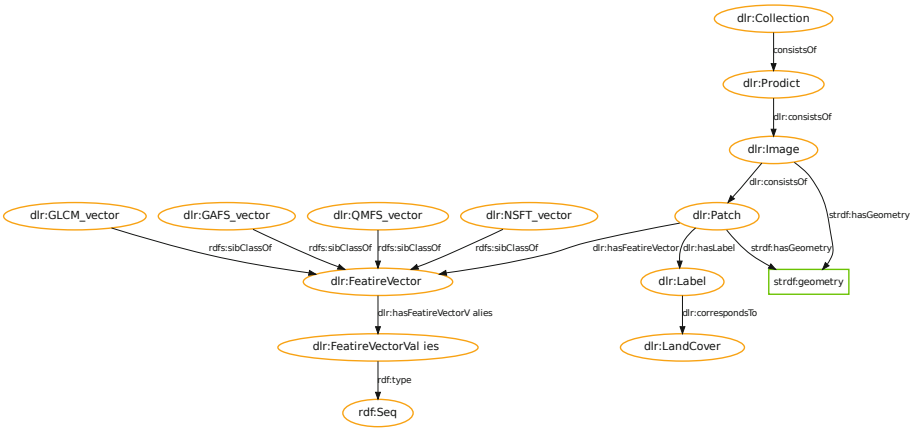


Fig. 4. A part of the DLR ontology for TerraSAR-X data

whereas the bottom level is more detailed (e.g., bridge, roads, river, channel, etc.). We expect that it will be further enriched later in the project when more TerraSAR-X images would have been processed, and our knowledge discovery techniques might be able to recognize a greater number of semantic classes.

A part of the class hierarchy of the DLR ontology is shown in Figure 4. The data property `strdf:hasGeometry` is also shown to give the reader an understanding of where geospatial information lies. It was noted in the beginning of Section 3 that a TerraSAR-X image has a spatial extent specified using WGS84 coordinates. We use these coordinates to construct a geometry in polygon format

projected to the WGS84 reference system for the whole image. The geometry is specified in Well-Known Text (WKT) format using the constructs available in stSPARQL as explained earlier, in Section 4. We also construct a geometry in polygon format projected to the WGS84 reference system for each patch of an image, because it would be infeasible to derive it using a SPARQL query with a variable binding. The geometry is needed because we would also want to compare patches between different images, which demands taking global position of the patch into account.

Some stRDF triples that have been produced from an actual product of the provided dataset are shown below so the reader can appreciate the kinds of data that are generated as instances of the classes of the ontology:

```
dlr:Product_1 rdf:type dlr:Product .
dlr:Product_1 dlr:hasImage dlr:Image_1.tif .
dlr:Product_1 dlr:hasName "TSX1_SAR"^^xsd:string .
dlr:Product_1 dlr:hasXMLfilename "TSX1_SAR.xml"^^xsd:string .
dlr:Image_1.tif rdf:type dlr:Image .
dlr:Image_1.tif dlr:hasName "IMAGE_HH_SRA_spot_047.tif"^^xsd:string .
dlr:Image_1.tif dlr:consistsOf dlr:Patch_1.jpg .
dlr:Image_1.tif strdf:hasGeometry "POLYGON ((12 45, 13 45, 13 46, 12 46,
                                                    12 45))"^^strdf:WKT .
dlr:Patch_1.jpg rdf:type dlr:Patch .
dlr:Patch_1.jpg dlr:hasName "Patch_200_0_0.jpg"^^xsd:string .
dlr:Patch_1.jpg dlr:hasSize "200"^^xsd:int .
dlr:Patch_1.jpg dlr:hasIndexI "0"^^xsd:int .
dlr:Patch_1.jpg dlr:hasIndexJ "0"^^xsd:int .
dlr:Patch_1.jpg strdf:hasGeometry "POLYGON ((12 44, 13 44, 13 45, 12 45,
                                                    12 44))"^^strdf:WKT .
dlr:Patch_1.jpg dlr:hasGAFS_vector dlr:GAFS_2_2_1 .
dlr:Patch_1.jpg dlr:hasLabel dlr:Label_1 . dlr:Label_1 rdf:type dlr:Label.
dlr:Label_1 dlr:correspondsTo dlr:Bridge .
dlr:GAFS_2_2_1 rdf:type dlr:GAFS_Vector .
dlr:GAFS_2_2_1 dlr:hasFeatureVectorValues dlr:GAFS_2_2_1_values .
```

5.2 Queries in the Virtual Earth Observatory

The purpose of this section is to show that with the work carried out in TELEIOS we significantly improve the state-of-art in EO portals such as EOWEB-NG that are aimed at end-user querying, but also data management systems available in EO data centers today, such as DIMS [28]. We first present a categorization of queries that are possible in the Virtual Earth Observatory. Then, we show how some of these queries can be expressed using the query language stSPARQL. In the Virtual Earth Observatory prototype that is currently under development in TELEIOS, the system Strabon is used for the storage of stRDF data while a visual query builder, presented in [16], is used to allow end-users to pose queries easily.

The following are some classes of queries that can be expressed by users of the Virtual Earth Observatory. The categorization presented is not exhaustive, but serves to illustrate the expressive power of our annotation schemes and the query language stSPARQL.

1. *Query for a product and its metadata.* This type of query is based on the metadata extracted from the XML file of the TerraSAR-X products (e.g., time and area of acquisition, sensor, imaging mode, incidence angle).

2. *Query for an image and its metadata.* This type of query is based on the image and its attached metadata (e.g., geographic latitude/longitude).
3. *Query for images of products that contain patches that have certain properties.* Queries of this type can be further categorized as follows:
   (a) *Query by the land cover/use class of a certain patch.* This type of query is based on the annotations of the patches, according to the land cover/use classification scheme presented in [6].
   (b) *Query by the land cover/use class of a patch and the qualitative or quantitative spatial properties of a patch.* This type of query allows us to query for patches with some land cover/use class that are spatially related to other patches or to a user defined area. Here one can use various qualitative or quantitative spatial relations (e.g., topological, cardinal directions, orientation, distance) [21,23] .
   (c) *Query by correlating the land cover/use class of more than one patch that have various qualitative or quantitative spatial relations between them.* This type of query extends the previous query by allowing the correlation based on land cover/use class of multiple patches with various spatial relations between them.
   (d) *Query that involves features of a patch but also other properties like the land cover/use class and spatial relations.* This type of query is based on the parameters of the feature extraction algorithms discussed in Section 3. Using feature values in queries is very useful when we want to distinguish patches of the same semantic class that differ on specific properties.

By examining the above types of queries, we see that existing EO portals such as EOWEB-NG and DIMS offer partial or full support for asking queries of type 1 and 2, but cannot be used to answer any of the queries of type 3 and its subcategories. These are queries that can only be asked and answered if the knowledge discovery techniques of Section 3 are applied to TerraSAR-X images and relevant knowledge is extracted and captured by semantic annotations expressed in stRDF. In other words these queries are made possible for users due to the advances of TELEIOS technologies.

We procceed with examples of queries from the above classes:

– *Class 3(c).* Find all patches containing water limited on the north by a port, at a distance of no more than 200 meters.

```
SELECT  ?PA ?PGEO1
WHERE { ?IM dlr:consistsOf ?PA1 . ?PA1 rdf:type dlr:Patch .
        ?IM dlr:consistsOf ?PA2 . ?PA2 rdf:type dlr:Patch .
        ?PA1 strdf:hasGeometry ?PGEO1 . ?PA1 dlr:hasLabel ?LA1 .
        ?LA1 rdf:type dlr:Label . ?LA1 dlr:correspondsTo dlr:Water .
        ?PA2 strdf:hasGeometry ?PGEO2 . ?PA2 dlr:hasLabel ?LA2 .
        ?LA2 rdf:type dlr:Label . ?LA2 dlr:correspondsTo dlr:Port .
        FILTER (strdf:above(?PGEO1,?PGEO2) &&
                strdf:contains(strdf:buffer(?PGEO2,0.005),?PGEO1)) . }
```

The results of this type of query are presented in Figure 5. Such a result can be useful for a port authority in order to monitor the port area. For other

Fig. 5. Query results projected on Google Maps

sites, this query can be extended in order to improve navigational safety in coastal regions near ports and other marine terminals that experience heavy traffic by large crude-oil carriers, towed barges, and other vessels of deep draught or restricted manoeuvrability.

– *Class 3(d).* Find all patches that correspond to a bridge, according to the 1^{st} feature value of the Gabor algorithm with 4 scales and 6 orientations.

```
SELECT  ?PA ?PGEO
WHERE { ?PA rdf:type dlr:Patch .
        ?PA dlr:hasLabel ?L . ?PA dlr:hasGeometry ?PGEO .
        ?L rdf:type dlr:Label . ?L dlr:correspondsTo dlr:Bridge .
        ?PA dlr:hasGAFSvector ?V . ?V rdf:type dlr:GAFS_Vector .
        ?V dlr:hasScales 4 . ?V dlr:hasOrientations 6 .
        ?V dlr:hasFeatureVectorValues ?FV . ?FV rdf:_1 ?FV1 .
        FILTER (5.0 < ?FV1 && ?FV1 < 35.0) . }
```

This type of query can be extended from only one feature value to the full dimension of the feature vector (e.g., for all 48 feature values of the Gabor algorithm with 4 scales and 6 orientations) and it can be used to distinguish objects of the same semantic class that differ on specific values. This query can be useful for EO scientists in order to extract new knowledge from feature vector values. For example, one can use such a query to identify certain values that form a property for the object in question (e.g., orientation). We expect the final version of our Virtual Earth Observatory to allow only users belonging to a particular role (e.g., scientists inside DLR) to execute such queries but we have not so far implemented this functionality.

– *Class 3.* Find all patches containing seagrass detritus and algae on shores that are identified as recreational beaches.
The results of this type of query can be useful for coastal management to seek to retain seagrass meadows and also to ensure that seagrass detritus stays on the beach and in the water. In Europe the desire for a clean beach has been taken to the point of daily raking and removal of algae for amenity

reasons; some fashionable beaches in France are even perfumed. However, such sanitisation has its costs, reporting a local loss of seabirds following the commencement of raking [8].

6 Related Work

TELEIOS is a multidisciplinary research effort bringing together contributions from database management, semantic web, remote sensing and knowledge discovery from satellite images. We now review some of the most relevant research efforts in these areas, and compare them with the work carried out in TELEIOS which has been presented in this paper.

In the context of the Semantic Web, the development of geospatial extensions to SPARQL has received some attention recently which resulted in the creation of a forthcoming OGC standard for querying geospatial data encoded in RDF, called GeoSPARQL [18]. GeoSPARQL draws on the concepts developed in earlier languages such as SPARQL-ST [19], SPAUK [9] and the original version of stSPARQL [12].

There have been some works in the past where ontologies have been applied to the modeling of EO data [20,4] or in a similar virtual observatory context [22,17]. TELEIOS has benefited from the modeling concepts developed in these efforts and has tried to reuse parts of these public ontologies whenever possible.

Finally, the vision of having knowledge discovery and data mining from satellite images as a fundamental capability of information systems for today's EO data centers has been stressed in earlier project KEO/KIM[16] funded by the European Space Agency, and the US project GeoIRIS [24]. Compared to these projects, TELEIOS has a much stronger technical foundation because it builds on state of the art database and semantic web technologies, as well as more advanced knowledge discovery and data mining techniques.

7 Conclusions

In this paper we report on a Virtual Earth Observatory that we are currently building in the context of the European project TELEIOS. Given the rapidly growing EO data archives, TELEIOS addresses the need for scalable access to petabytes of EO data and the discovery of knowledge that can be used in applications. The main focus is on knowledge discovery from EO images and geospatial Semantic Web technologies (stRDF and stSPARQL). We discuss in detail how the developed technologies can be deployed to improve the state-of-art in EO portals by enabling queries that capture the semantics of the content of the images. The work presented reflects what we have achieved in the first one and a half years of the project. Future work includes enrichment of the DLR ontology and integration of a visual query builder, presented in [16].

[16] `http://earth.esa.int/rtd/Projects/KEO/index.html`

References

1. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semantic Web Inf. Syst. (2009)
3. Bratasanu, D., Nedelcu, I., Datcu, M.: Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications. IEEE JSTARS (2011)
4. Carlino, C., Elia, S.D., Vecchia, A.D., Iacovella, M., Iapaolo, M., Scalzo, C.M., Verdino, F.: On sharing Earth Observation concepts via ontology. In: ESA-EUSC (2008)
5. Datcu, M., D'Elia, S., King, R., Bruzzone, L.: Introduction to the special section on image information mining for earth observation data. IEEE Transactions on Geoscience and Remote Sensing 45(4), 795–798 (2007)
6. Dumitru, C.O., Datcu, M., Koubarakis, M., Sioutis, M., Nikolaou, B.: Ontologies for the VO for TerraSAR-X data. Del. D6.2.1, FP7 project TELEIOS (2012)
7. Dumitru, C.O., Molina, D.E., Cui, S., Singh, J., Quartulli, M., Datcu, M.: KDD concepts and methods proposal: report & design recommendations. Del. 3.1, FP7 project TELEIOS (2011)
8. Harvey, N., Caton, B.: Coastal Management in Australia. Meridian Series. Oxford University Press (2003)
9. Kolas, D., Self, T.: Spatially-Augmented Knowledgebase. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 792–801. Springer, Heidelberg (2007)
10. Koubarakis, M., Karpathiotakis, M., Kyzirakos, K., Nikolaou, C., Vassos, S., Garbis, G., Sioutis, M., Bereta, K., Manegold, S., Kersten, M., Ivanova, M., Pirk, H., Zhang, Y., Kontoes, C., Papoutsis, I., Herekakis, T., Michail, D., Datcu, M., Schwarz, G., Dumitru, O.C., Molina, D.E., Molch, K., Giammatteo, U.D., Sagona, M., Perelli, S., Klien, E., Reitz, T., Gregor, R.: Building Virtual Earth Observatories using Ontologies and Linked Geospatial Data. In: Proceedings of the 6*th* International Conference on Web Reasoning and Rule Systems (Technical Communication), Vienna, Austria, September 03-08 (2012)
11. Koubarakis, M., Karpathiotakis, M., Kyzirakos, K., Nikolaou, C., Sioutis, M.: Data Models and Query Languages for Linked Geospatial Data. Invited papers from 8th Reasoning Web Summer School, Vienna, Austria, September 03-08 (2012)
12. Koubarakis, M., Kyzirakos, K.: Modeling and Querying Metadata in the Semantic Sensor Web: The Model stRDF and the Query Language stSPARQL. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 425–439. Springer, Heidelberg (2010)
13. Koubarakis, M., Kyzirakos, K., Karpathiotakis, M.: Data Models, Query Languages, Implemented Systems and Applications of Linked Geospatial Data. Tutorial Presented at ESWC (2012), `http://www.strabon.di.uoa.gr/tutorial-eswc-2012`

14. Koubarakis, M., Kyzirakos, K., Karpathiotakis, M., Nikolaou, C., Vassos, S., Garbis, G., Sioutis, M., Mpereta, C., Michail, D., Kontoes, C., Papoutsis, I., Herekakis, T., Manegold, S., Kersten, M., Ivanova, M., Pirk, H., Datcu, M., Schwarz, G., Dumitru, O.C., Molina, D.E., Molch, K., Giammatteo, U.D., Sagona, M., Perelli, S., Reitz, T., Klien, E., Gregor, R.: Building earth observatories using semantic web and scientific database technologies. In: Proceedings of the 38*th* International Conference on Very Large Databases (Demo Paper), Istanbul, Turkey, August 27-31 (2012)
15. Koubarakis, M., Kyzirakos, K., Nikolaou, B., Sioutis, M., Vassos, S.: A data model and query language for an extension of RDF with time and space. Del. 2.1, FP7 project TELEIOS (2011)
16. Sagona, M., Di Giammatteo, U., Gregor, R., Klien, E.: The TELEIOS infrastructure - version I. Del. D1.3, FP7 project TELEIOS (2012)
17. McGuinness, D.L., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J.L., Middleton, D.: The virtual solar-terrestrial observatory: A deployed semantic web application case study for scientific research. In: AAAI (2007)
18. Open Geospatial Consortium Inc: GeoSPARQL - A geographic query language for RDF data (November 2010)
19. Perry, M.: A Framework to Support Spatial, Temporal and Thematic Analytics over Semantic Web Data. Ph.D. thesis, Wright State University (2008)
20. Podwyszynski, M.: Knowledge-based search for Earth Observation products. Master's thesis, Passau Univ. (2009)
21. Randell, D.A., Cui, Z., Cohn, A.G.: A Spatial Logic based on Regions and Connection. In: Proceedings of the 3*rd* International Conference on Knowledge Representation and Reasoning (1992)
22. Raskin, R., Pan, M.: Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences (2005)
23. Renz, J., Nebel, B.: Qualitative spatial reasoning using constraint calculi. In: Handbook of Spatial Logics, pp. 161–215. Springer (2007)
24. Shyu, C.R., Klaric, M., Scott, G., Barb, A., Davis, C., Palaniappan, K.: GeoIRIS: Geospatial Information Retrieval and Indexing System - Content Mining, Semantics Modeling, and Complex Queries. In: IEEE TGRS (2007)
25. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12), 1349–1380 (2000)
26. Snoek, C.G.M., Smeulders, A.W.M.: Visual-concept search solved? IEEE Computer 43(6), 76–78 (2010)
27. Tzouveli, P., Simou, N., Stamou, G., Kollias, S.: Semantic classification of byzantine icons. IEEE Intelligent Systems 24, 35–43 (2009)
28. Wolfmüller, M., Dietrich, D., Sireteanu, E., Kiemle, S., Mikusch, E., Böttcher, M.: Data Flow and Workflow Organization - The Data Management for the TerraSAR-X Payload Ground Segment. In: IEEE TGRS (2009)