BUILDING VIRTUAL EARTH OBSERVATORIES USING ONTOLOGIES, LINKED GEOSPATIAL DATA AND KNOWLEDGE DISCOVERY ALGORITHMS

M. Koubarakis, M. Sioutis, K. Kyzirakos, M. Karpathiotakis, C. Nikolaou, S.Vassos, G. Garbis, K. Bereta

National and Kapodistrian University of Athens, Greece

M. Datcu, G. Schwarz, O.C. Dumitru, D.E. Molina, K. Molch

German Aerospace Center (DLR), Germany

ABSTRACT

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 knowledge discovery framework of TELEIOS and discuss how it can be used together with ontologies and linked geospatial data for the development of a Virtual Earth Observatory for TerraSAR-X 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 the last few years, and it is currently reaching petabytes (PBs) in many satellite archives. 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. TELEIOS (`http://www.earthobservatory.eu/`) 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.

The contributions of this paper are the following: (i) we present the knowledge discovery framework developed in TELEIOS and give details of its application to radar images captured by TerraSAR-X, the synthetic aperture radar (SAR) satellite of TELEIOS partner German Aerospace Data Center (DLR), (ii) we present briefly the use of data model stRDF and its query language stSPARQL in the construction of Virtual Earth Observatories, (iii) 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. A first prototype of the TELEIOS Virtual Earth Observatory architecture is also demonstrated in [1], using a forest fire monitoring application as example.

THIS WORK HAS BEEN FUNDED BY THE FP7 PROJECT TELEIOS (257662).

2. 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.

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.

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 SAR 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 also, identifying objects and features in them is challenging. For example, mining of EO images based on content analysis is very different from mining images of faces or animals, due to the nature of actual features (e.g., eyes, ears, stripes, or wings) that have known relationships and therefore promote the differentiation of classes. Moreover, in SAR images that are studied in TELEIOS, additional problems arise from the fact that these images have different acquisition properties than optical images. Essentially, although SAR products may look like optical images, they are mathematical products that rely on delicate radar measurements.

In [2] we presented a detailed analysis of TerraSAR-X Level 1b products and identified the ones that we will use for our knowledge discovery research and the Virtual Earth

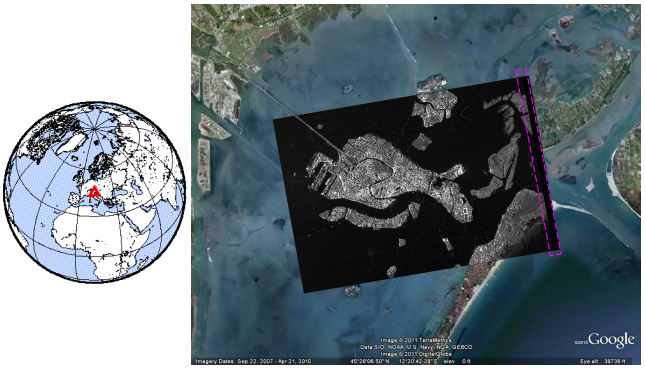

Fig. 1. Overlay on Google Earth and location of the Venice site

Observatory implementation in TELEIOS. Each TerraSAR-X product comprises a XML file which defines in detail the data types, valid entries, and allowed attributes of a product, and a TerraSAR-X image. Additionally, a preview of the product in GeoTIFF[1] format is given as a quick-look image georeferenced to the WGS84 coordinate reference system, and annotated with latitude/longitude coordinates. Figure 1 shows an example of a quick-look image of Venice projected on Google Earth and its position on the globe. The quick-look image serves only as a preview of the TerraSAR-X product and is not to be mistaken with the actual TerraSAR-X 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, 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 [2].

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 together with the number and kind of features they produce are presented in detail in [2].

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. [2] presents detailed experimental results that have been obtained by applying our techniques to TerraSAR-X images to detect the 35 classes presented in [3, chap. 3]. The semantic class labels are concepts from an RDFS ontology, presented in Section 4, which we have defined especially for the Virtual Earth Observatory for TerraSAR-X data.

3. 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. 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 [4].

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 query stRDF data and enable the development of EO applications like the Virtual Earth Observatory for TerraSAR-X data described in the following section. For example, one can use stSPARQL to express in a single query an information request such as the following: “Find images containing ports

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

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 semantics of the content of the products are not included in the archived metadata, thus they cannot be used as search criteria. In EOWEB-NG and other similar Web interfaces, search criteria consist solely of a hierarchical organization of available products (e.g., high resolution optical data, Synthetic Aperture Radar data, their subcategories) combined with a temporal and geographic selection menu.

The stRDF model and stSPARQL query language have been implemented in the system Strabon [4] which is publicly available as open source software (`http://www.strabon.di.uoa.gr/`).

4. 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.

The first step in the development of the Virtual Observatory for TerraSAR-X data was the construction of an RDFS ontology, the DLR ontology (`http://www.earthobservatory.eu/ontologies/dlrOntology.owl`), which captures the contents of the Virtual Earth Observatory. The DLR ontology comprises the following major parts: (i) the part that captures the hierarchical structure of a product and the XML metadata associated with it (e.g., time and area of acquisition, sensor, imaging mode, incidence angle), (ii) the part that defines the RDFS classes and properties that formalize the outputs of the knowledge discovery step (e.g., patch, feature vector), (iii) the part that defines the land cover/use classification scheme [3] for annotating image patches that was constructed while experimenting with the knowledge discovery framework presented in Section 2.

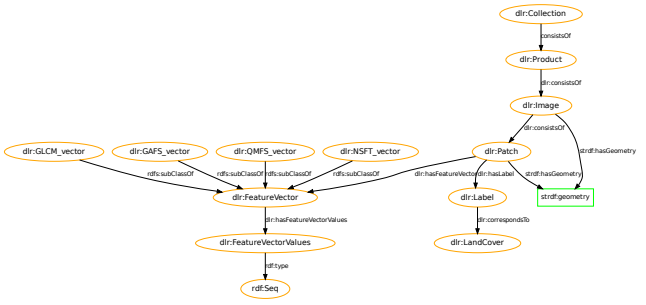

Fig. 2. A part of the DLR ontology for TerraSAR-X data

A part of the class hierarchy of the DLR ontology is shown in Figure 2. The data property `strdf:geometry` is also shown to give the reader an understanding of where geospatial information lies. It was noted in the beginning of Section 2 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 3. 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.

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 [3].

   (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) [5, 6] .

   (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 2. 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 2 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 proceed with examples of queries of type 3:

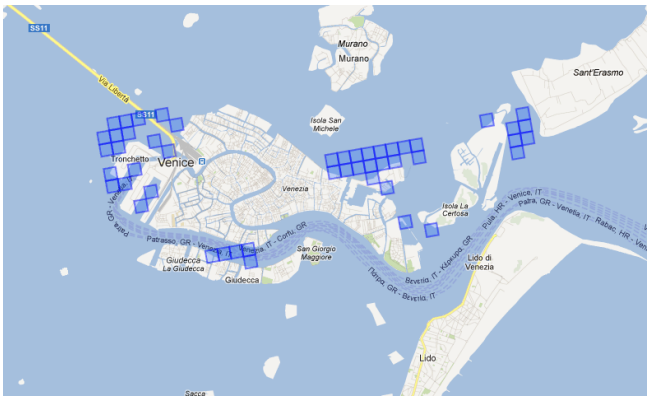

Fig. 3. Query results projected on Google Maps

- *Find all patches containing water limited on the north by a port.*

```
SELECT  ?PA1 ?PGEO1
WHERE {
   ?PA1 rdf:type dlr:Patch .
   ?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(?PGEO2,0.05),
                              ?PGEO1)).
}
```

The results of this type of query are presented in Figure 3. Such a result can be useful for a port authority in order to monitor the port area. For other 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.

- *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. However, such sanitisation has its costs, reporting a local loss of seabirds following the commencement of raking.

5. CONCLUSIONS

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 [7, Appendix A].

6. REFERENCES

[1] M. Koubarakis, K. Kyzirakos, M. Karpathiotakis, C. Nikolaou, S. Vassos, G. Garbis, M. Sioutis, K. Bereta, D. Michail, C. Kontoes, I. Papoutsis, T. Herekakis, S. Manegold, M. L. Kersten, M. Ivanova, H. Pirk, Y. Zhang, M. Datcu, G. Schwarz, C. Dumitru, D. Espinoza-Molina, K. Molch, U. Di Giammatteo, M. Sagona, S. Perelli, T. Reitz, E. Klien, and R. Gregor, "TELEIOS: A Database-Powered Virtual Earth Observatory," *PVLDB*, vol. 5, no. 12, pp. 2010–2013, 2012, demo paper.

[2] C. O. Dumitru, D. Espinoza-Molina, S. Cui, J. Singh, M. Quartulli, and M. Datcu, "KDD concepts and methods proposal: report & design recommendations," Del. 3.1, TELEIOS project, 2011.

[3] C. O. Dumitru, M. Datcu, M. Koubarakis, M. Sioutis, and C. Nikolaou, "Ontologies for the VO for TerraSAR-X data," Del. 6.2.1, TELEIOS project, 2012.

[4] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis, "Strabon: A Semantic Geospatial DBMS," in *Proceedings of the 11th International Semantic Web Conference*, 2012.

[5] D. A. Randell, Z. Cui, and A. G. Cohn, "A Spatial Logic based on Regions and Connection," in *KR*, 1992.

[6] J. Renz and B. Nebel, "Qualitative spatial reasoning using constraint calculi," in *Handbook of Spatial Logics*, pp. 161–215. 2007.

[7] M. Sagona, U. D. Giammatteo, R. Gregor, and E. Klien, "The TELEIOS infrastructure - version I," Del. 1.3, TELEIOS project, 2012.