BUILDING REMOTE SENSING APPLICATIONS USING SCIENTIFIC DATABASE AND SEMANTIC WEB TECHNOLOGIES

The TELEIOS team[*]

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 is hidden in them. TELEIOS builds on scientific database technologies (array databases, SciQL, data vaults) and Semantic Web technologies (stRDF and stSPARQL) implemented on top of a state of the art column store database system (MonetDB). In this paper we outline the gains that Earth Observation organizations can have from these technologies by presenting a detailed example of a fire monitoring service that we have completed.

1. INTRODUCTION

TELEIOS (`http://www.earthobservatory.eu/`) is a European project that addresses the need for scalable access to petabytes of Earth Observation (EO) data and the effective discovery of knowledge hidden in it. In the first twenty two months of the project, we have developed state-of-the-art techniques in Scientific Databases and Semantic Web, and we have applied them to the management of EO data.

In this paper we give a detailed example of a fire monitoring service that we have just completed using TELEIOS technologies for the National Observatory of Athens (NOA). In this way, we outline the vision of TELEIOS and explain in detail why it goes beyond operational systems for this application currently deployed in various EO organizations. The vision and contributions of TELEIOS are also illustrated in [1] where a first prototype of the TELEIOS Virtual Earth Observatory is demonstrated.

TELEIOS is unique among similar EO projects because it bases its innovation to the state of the art on its original contributions to data models, query languages and other database techniques for EO. The first such query language is SciQL, a new SQL-based query language for scientific applications which provides efficient array manipulation primitives [2]. SciQL is used to perform low level image processing and image content analysis in a high-level declarative way. It has been implemented on top of the pioneer column-store database system MonetDB (`http://www.monetdb.org/`) which has many of the capabilities we need for scalable querying of petabytes of satellite image data. Second, the data vault technique is used to enable the efficient access to large archives of image data and metadata in a fully transparent way, independently of their format, size and location [3]. Finally, stRDF (a geospatial extension of RDF) is used to represent satellite image metadata, knowledge extracted from satellite images and auxiliary geospatial datasets encoded as linked data. The query language stSPARQL can then be used for interacting with these high level, semantic representations of EO data effectively to enable the easy development of applications. stSPARQL has been implemented in the semantic geospatial DBMS Strabon (`http://www.strabon.di.uoa.gr`) which utilizes MonetDB as a back end and has been shown to scale to billions triples [4].

2. THE FIRE MONITORING APPLICATION OF NOA

NOA has been archiving and processing on a routine basis large volumes of satellite images of different spectral and spatial resolutions in combination with auxiliary geoinformation layers (e.g., land use/land cover data, administrative boundaries) to generate, validate and deliver fire-related products. In this context NOA has been developing a real-time fire hotspot detection service for effectively monitoring a fire-front. Since 2007, NOA operates an MSG/SEVIRI acquisition station, and has been systematically archiving raw satellite images on a 5 and 15 minutes basis, the respective temporal resolutions of MSG-1 and MSG-2. The archives of raw imagery are now in the order of 2 TB, corresponding to the summer fire periods of the last five years.

The fire monitoring service active in NOA *before* TELEIOS (presented in Figure 1) can be summarized as follows.

First, the ground-based receiving antenna collects all spectral bands from MSG-1 and MSG-2 every 5 and 15 minutes respectively. Then, the raw datasets are decoded and temporarily stored in the METEOSAT Ground Station as wavelet compressed images. Finally, an application, written

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

*M. Koubarakis, K. Kyzirakos, M. Karpathiotakis, C. Nikolaou, S. Vassos, G. Garbis, M. Sioutis, K. Bereta (University of Athens), C. Kontoes, I. Papoutsis, T. Herekakis (National Observatory of Athens), D. Michail (Harokopio University of Athens), S. Manegold, M. Kersten, M. Ivanova, H. Pirk, Y. Zhang (Centrum Wiskunde & Informatica), M. Datcu, G. Schwarz, O. C. Dumitru, D. E. Molina, K. Molch (German Aerospace Center), U. D. Giammatteo, M. Sagona, S. Perelli (Advanced Computer Systems), T. Reitz, E. Klien, R. Gregor (Fraunhofer Inst. for Computer Graphics Research).

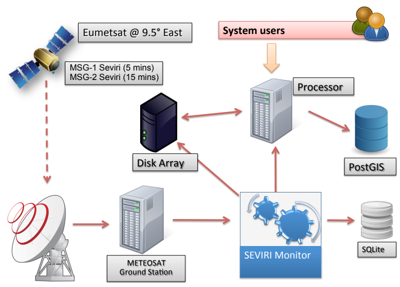

Fig. 1. The before TELEIOS NOA fire monitoring service

in Python, manages the data stream in real-time by offering the following functionality:

1. Extract and store the raw file metadata (e.g., sensor type, acquisition time) in an SQLite database. Such a step is required as one image is comprised of multiple raw files which might arrive out-of-order.

2. Filter the raw data files, disregarding non-applicable data for the fire monitoring scenario, and dispatch them to a dedicated disk array for permanent storage.

3. Remotely trigger the processing chain by transferring the appropriate spectral bands via FTP to a dedicated machine and initiating the distinct processing steps described in [5]. These steps are: (i) cropping the image, (ii) georeferencing to Hellenic Geodetic Reference System 1987 (HGRS87), (iii) classifying the image pixels as "fire" or "non-fire" using the algorithm of [6], and finally (iv) exporting the final product to raster and vector formats.

4. Dispatch the derived products to the disk array and additionally store them to a PostGIS database system.

The products that are stored in PostGIS cover the geographical area of Greece and are disseminated to the end user community through a web application.

One of the goals of TELEIOS is to improve the hotspot detection and the fire monitoring service of NOA described above. The main issues that need to be addressed are the following.

The thematic accuracy of the generated products has to be refined in a clear and systematic way, to ensure the reliability and transferability of the service to other geographic areas. The main problem with the current thematic accuracy is the existence of false alarms and omission errors in the fire detection technique that relate to the following scenarios.

- Cases of hotspots occurring in the sea or in locations represented by fully inconsistent land use/land cover classes, like urban or permanent agriculture areas. If these hotspots correspond to real fires, these fires occur in the vicinity of coasts or urban areas, but due to the low spatial pixel resolution of the MSG/SEVIRI instrument and errors in image georeferencing, the hotspots wrongly appear to be over inconsistent underlying land use/land cover classes. This type of error could be easily corrected if derived hotspot products are compared with auxiliary GIS layers by a NOA operator. However, this would certainly require time for manual GIS layer integration and visual interpretation, an operation that is not possible in the available 5 minute time frame.

- Cases of hotspots located outside forested areas. These can be false fire detections due to known problems with existing hotspot detection algorithms (e.g., inappropriate fire/no-fire thresholds in the algorithm of [6]). They can also be fires located in big agricultural plains that are put by farmers as part of their agricultural practices. Whichever the case, they are not real forest fires, and they are not emergency situations to be handled. This type of noisy information could be avoided if derived hotspot products are combined together with land use/land cover information, again an operation that cannot be done manually in the 5 minute time frame.

- Spatial and temporal inconsistencies in the final product. Today hotspot detection at a given time is done by using a single image acquisition corresponding to that time, without taking into consideration hotspots and their locations in previous image acquisitions. Given the inaccuracies of existing hotspot detection algorithms [6], this single-scene processing approach results in some spatial and temporal inconsistencies between the different observations. A simple heuristic, which would result in significant noise removal, is to check the number of times a specific fire was detected over the same or near the same geographic location during the last hour(s), considering the observation's temporal and spatial persistence, and hence attributing a level of confidence to each detected pixel.

There is also the need to generate added-value thematic maps combining diverse information sources. As a service provider NOA aims at delivering to the end user community reliable and comprehensive information for fire related emergency situations. Although vector shapefiles are useful for analysis in the aftermath of a crisis, in real-time emergency response scenarios, civil protection agencies and local firefighting teams find it more useful to refer to a map depicting the active fire-front and its evolution in the last hours/days and identify nearby crucial infrastructure (e.g., hospitals, fire hydrants). This is of paramount importance for the effective allocation of resources during the crisis. Therefore, a desired functionality that is currently missing is automatic map generation enriched with easily accessible geoinformation layers.

Finally, the dispersion of the various processes of the fire monitoring service in many machines and pieces of software

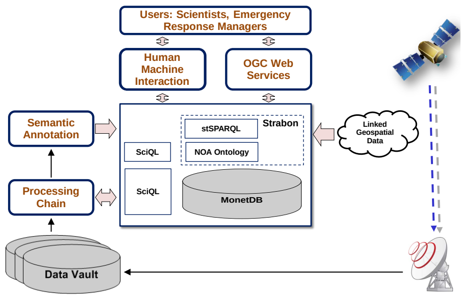

Fig. 2. The improved NOA fire monitoring service

makes it difficult for NOA to keep all functionalities synchronized. There is no consistent management policy, but various independent components (as seen in Figure 1). This in not a good solution for effectively managing the raw satellite imagery, the generated products and the static GIS layers. A more robust and user-friendly management system is needed that will allow the integration and customization of the available capacities.

3. USING TELEIOS TECHNOLOGIES IN NOA

Let us now describe the implementation of the fire monitoring service of NOA using TELEIOS technologies (shown in Figure 2) and point out the gains that NOA and other EO organizations can have from these technologies, not only for fire monitoring, but for many other EO applications.

Loading. One of the major issues that arise when dealing with EO data is the abundance of available file formats. The use of an external program that transforms the original satellite image format (HRIT) into a table/array representation is a major hurdle, not only in terms of inconvenience for the user, but also in terms of performance. A generic solution for this problem, developed in TELEIOS is the *data vault*.

The main idea of the data vault is to make the DBMS aware of external file formats and keep the knowledge how to convert data from external file formats into tables or arrays inside the database. With this, inserting external files (of known format) into the database basically consists of copying the files "as-is" into a directory that is under exclusive control of the database. Only after issuing queries that actually access data of a certain file, the DBMS will take care of loading the data from the file into the respective table or array [3].

Cropping and georeference. NOA is interested only in a specific part of the received image. Cropping the image early on, significantly reduces the input size of the remaining image processing operations and thus the time required for the execution of the processing chain. After the cropping operation, the image is georeferenced to HGRS87. Cropping, georeferencing and other low-level image processing operations are implemented in TELEIOS using *database techniques*. The main contribution of TELEIOS in this area is the development of the query processing language SciQL.

SciQL is a new SQL-based query language for scientific applications with arrays as first-class citizens. SciQL uses multi-dimensional arrays to represent EO data. This allows us to store EO data (e.g., satellite images) in the database, and query and manipulate their content transparently within the high-level declarative database query language. This has three important advantages. First, it allows us to express low level image processing (e.g., cropping, georeferencing) and image content analysis (e.g., feature extraction, pixel classification) in a user-friendly high-level declarative language that provides efficient array manipulation primitives. Second, it opens up these algorithms to be optimized by the (extended) query optimizer of the DBMS. Third, using the seamless integration and symbiosis of relational tables and arrays, query processing and knowledge discovery can exploit both image metadata and image data at the same time [2].

Classification. The fire classification module of the processing chain receives as input the cropped and georeferenced image with two pixel temperatures, each derived from the IR bands 3.9 and 10.8. The algorithm [6] uses the structural grouping capabilities of SciQL, in order to gather for each pixel the values of its neighbors inside a 3x3 window and computes the standard deviation of the temperatures inside the window. Afterwards, it classifies each pixel as "fire", "potential fire" or "no fire". A set of 4 thresholds, one for the temperature of the IR 3.9 band, one for the difference between the temperatures of the IR 3.9 and the IR 10.8 band, and two for the standard deviations of the two temperatures, are used for the classification of the pixel.

Final products. Finally, a SciQL query selects the pixels which were classified as "fire" or "potential fire" and outputs an ESRI shapefile that contains the POLYGON description of each pixel and a confidence level of 0.5 or 1, for "fire" and "potential fire" respectively.

Combining final products with auxiliary data. As described in Section 2, the outputs of the hotspot detection processing chain need to be combined with auxiliary data to improve their thematic accuracy and to allow the automatic generation of added-value related maps. Both of these tasks are performed using *ontologies, linked geospatial data* and related technologies developed in TELEIOS.

stRDF is the first of these technologies. It is an extension of the W3C standard RDF, that enables 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 stRDF data. stRDF is used to encode image metadata, image content extracted using image analysis techniques and auxiliary GIS data using vocabulary from appropriate ontologies. stSPARQL is then used to query this stRDF data to enable the development of EO applications [4].

To be able to query NOA data using stSPARQL and combine them with linked data, the produced shapefiles and the initial raw data files are first transformed in RDF. Using RDF triples, each raw data file, ESRI shapefile or hotspot extracted from a shapefile is connected with the satellite and the sensor from which it is derived, as well as with the sensing date and time. Hotspots and shapefiles are also associated with the exact method (processing chain) which was used for their production and with the organization which is responsible for the production (e.g., NOA). Finally, hotspots are additionally connected with the region (pixel) where they lie and the confidence level (0.5 or 1) derived from the classification phase.

In order to enrich the dataset of NOA with auxiliary geospatial data we compiled in RDF the following datasets.

- *Corine Land Use/Land Cover* is a dataset of the European Environment Agency, that describes the environmental landscape of Europe, expressed in RDF.
- *Coastline of Greece* is a RDF dataset that describes the geometry of the coastline of Greece.
- *Greek Administrative Geography* is an ontology that describes the administrative divisions of Greece (prefecture, municipality, district etc.). The ontology has been populated with relevant data that are available in the Greek open government data portal.

Additionally, the dataset is enriched with the following data from the Linked Open Data Cloud.

- *OpenStreetMap* data expressed in RDF by the project LinkedGeoData.
- *GeoNames* is a gazetteer that collects both spatial and thematic information for various placenames around the world.

Improving thematic accuracy. The first important issue in the fire monitoring service of NOA is automatically improving the accuracy of the detected hotspots. TELEIOS suggests stRDF as a common format for representing both derived hotspots, their metadata, and auxiliary layers, so that they can easily be combined in a single stSPARQL query. The thematic accuracy of the derived hotspots is now improved automatically by an additional process step that performs a series of stSPARQL update statements that update the RDF representation of the hotspots by taking into account relevant RDF datasets from the ones presented above. As an example, consider the following update operation.

```
DELETE {?h strdf:hasGeometry ?hGeo}
INSERT {?h strdf:hasGeometry ?dif}
WHERE {
  SELECT DISTINCT ?h ?hGeo
   (strdf:intersection(?hGeo, strdf:union(?cGeo)) AS ?dif)
  WHERE { ?h a noa:Hotspot; strdf:hasGeometry ?hGeo.
          ?c a coast:Coastline; strdf:hasGeometry ?cGeo.
          FILTER(strdf:anyInteract(?hGeo, ?cGeo))}
  GROUP BY ?h ?hGeo
  HAVING strdf:overlap(?hGeo, strdf:union(?cGeo))}
```

This operation, utilizing the dataset about the Greek coastline, retrieves hotspots that lie in the sea and deletes the part of their geometry that lies in the sea.

Improving automatic map generation. The automatic generation of fire maps enriched with relevant geoinformation is of paramount importance to NOA, since the creation of such maps in the past has been a manual process. Semantic Web technologies provide tools for handling heterogeneous data in a homogeneous way (stRDF/stSPARQL), while Linked Open Data Cloud supplies an abundance of data, like the datasets presented above, in addition to internal EO data. So, instead of manually combining heterogeneous data, a user can pose an stSPARQL query, using an stSPARQL endpoint, for each layer that she wants to depict in a map and overlay the retrieved data using the ability of Strabon to expose data in KML and GeoJSON.

4. CONCLUSIONS

Using a forest fire monitoring service as a representative example, we discussed how scientific database technologies (array databases, SciQL, data vaults) and Semantic Web technologies (stRDF and stSPARQL) can be deployed to support and improve the processing of large-scale EO data. While TELEIOS focuses only on remote sensing, the developed technologies can also be deployed in other scientific disciplines with similar data processing needs e.g., astronomy, meteorology, seismology, etc.

5. 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] Y. Zhang, M. L. Kersten, M. Ivanova, and N. Nes, "SciQL: bridging the gap between science and relational DBMS," in *IDEAS*, 2011, pp. 124–133.

[3] M. Ivanova, M. Kersten, and S. Manegold, "Data Vaults: A Symbiosis between Database Technology and Scientific File Repositories," in *Scientific and Statistical Database Management*, vol. 7338 of *LNCS*, pp. 485–494. Springer, 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] N. Sifakis, C. Iossifidis, C. Kontoes, and I. Keramitsoglou, "Wildfire Detection and Tracking over Greece Using MSG-SEVIRI Satellite Data," *Remote Sensing*, vol. 3, pp. 524–538, 2011.

[6] EUMETSAT, "Active fire monitoring with msg algorithm - theoretical basis document," *Report EUM/MET/REP/07/0170*, 2007, Eumetsat: Darmstadt, Germany.