TELEIOS: A Database-Powered Virtual Earth Observatory

[1] M. Koubarakis K. Kyzirakos M. Karpathiotakis C. Nikolaou S. Vassos
G. Garbis M. Sioutis K. Bereta [2] D. Michail [3] C. Kontoes I. Papoutsis
T. Herekakis [4] S. Manegold M. Kersten M. Ivanova H. Pirk Y. Zhang
[5] M. Datcu G. Schwarz O. C. Dumitru D. E. Molina K. Molch
[6] U. D. Giammatteo M. Sagona S. Perelli [7] T. Reitz E. Klien R. Gregor
[1] University of Athens [2] Harokopio University of Athens [3] National Observatory of Athens
[4] Centrum Wiskunde & Informatica [5] German Aerospace Center [6] Advanced Computer Systems
[7] Fraunhofer Inst. for Computer Graphics Research

ABSTRACT

TELEIOS is a recent European project that addresses the need for scalable access to petabytes of Earth Observation data and the discovery and exploitation 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). We demonstrate a first prototype of the TELEIOS Virtual Earth Observatory architecture, using a forest fire monitoring application as example.

1. INTRODUCTION

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 German Aerospace Center (DLR) is expected to reach 2 PB next year, while ESA estimates that it will be archiving 20 TB 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 huge.

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 in September 2010 and it will last for 3 years. In the first year of the project, we have started the development of state-of-the art techniques in Scientific Databases, Semantic Web and Image Information Mining, and have applied them to the management of EO data.

In [9], we have developed SciQL, a new SQL-based query language for scientific applications with arrays as first-class citizens. SciQL uses multi-dimensional arrays to represent EO data of various processing levels. 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, resampling, georeferencing etc.) as well as 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 DBMS's (extended) query optimizer. 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.

We have also developed the model stRDF, an extension of the W3C standard RDF, that allows the representation of geospatial data that changes over time [7, 5]. stRDF is accompanied by stSPARQL, an extension of the query language SPARQL 1.1 for querying stRDF data. stRDF and stSPARQL use OGC standards (Well-Known Text and Geography Markup Language) for the representation of temporal and geospatial data. Strabon is a fully implemented semantic geospatial database system that can be used to store linked geospatial data expressed in stRDF and query them using stSPARQL[2].

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 region is a water body) and auxiliary geospatial data sets encoded as linked data. One can then use stSPARQL to express in a single query an information request such as the following: "Find an image taken by a Meteosat second generation satellite on August 25, 2007 which covers the area of Peloponnese and contains hotspots corresponding to forest fires located within 2km from a major archaeological site". Encoding this information request today in a typical interface to an EO data archive such as EOWEB-NG[3] is

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

[2] http://www.strabon.di.uoa.gr

[3] http://eoweb.dlr.de/

Proceedings of the VLDB Endowment, Vol. X, No. Y

impossible, because domain-specific concepts such as "forest fires" are not included in the archive metadata, thus they cannot be used as search criteria. 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..

In [3, 4] 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 water-body, lake, forest, etc., or environmental monitoring ontologies with concepts such as forest fires, flood, etc.). These concepts are encoded in OWL ontologies and are used to annotate EO products. In this way, we attempt to close the semantic gap that exists between user requests and searchable information available explicitly in the archive.

But even if domain-specific concepts were included in the archive annotations, one would need to join them with information obtained from auxiliary data sources to answer the above query (e.g., Wikipedia to find the major archaeological sites in the Peloponnese, GeoNames to find their geographic location etc.). 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 be used to express information requests such as the above. The linked data web is being populated with geospatial data rapidly [1], thus we expect that languages such as stSPARQL (and the related forthcoming OGC standard GeoSPARQL [8]) will soon be mainstream extensions of SPARQL that can be used to access such data effectively.

The technologies developed in TELEIOS are implemented on top of the pioneer column-store database system MonetDB[4] and the geospatial RDF store Strabon.

In the remainder of this demo paper, we first present the basic concepts of the TELEIOS Virtual Earth Observatory. Then we present a high level view of the Earth Observatory architecture. Finally, we describe in more detail how we plan to demonstrate the Earth Observatory by giving a detailed example of a fire monitoring application that we have just completed using TELEIOS technologies.

2. BASIC CONCEPTS

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 satellite missions which is typically carried out as follows (Figure 1, grey part).

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 a variety of storage systems and are made available using a variety of policies depending on their volume and expected future use.

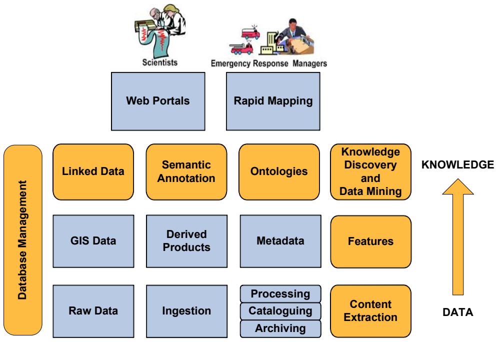


Figure 1: Concept view of the TELEIOS Virtual Earth Observatory

EO data centers also offer a variety of user services. For example, for scientists that want to utilize EO data in their research, the TELEIOS partner German Aerospace Center (DLR) offers the Web interface EOWEB-NG for searching, inspection and ordering of products. Space agencies such as DLR might also make various other services available aimed at specific classes of users. For example, the Center for Satellite Based Crisis Information (ZKI[5]) 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 the TELEIOS partner National Observatory of Athens (NOA) through its participation in the GMES SAFER program.

The TELEIOS advancements to today's state of the art in EO data processing are shown graphically with yellow color in Figure 1 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, burned area etc.) that characterize the content of an image [4].

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-like interface to EO data archives can be developed on top of a system like Strabon to enable end-users to pose very expressive queries. Rapid mapping applications can also take advantage of rich seman-

[4] http://www.monetdb.org/

[5] http://www.zki.dlr.de/

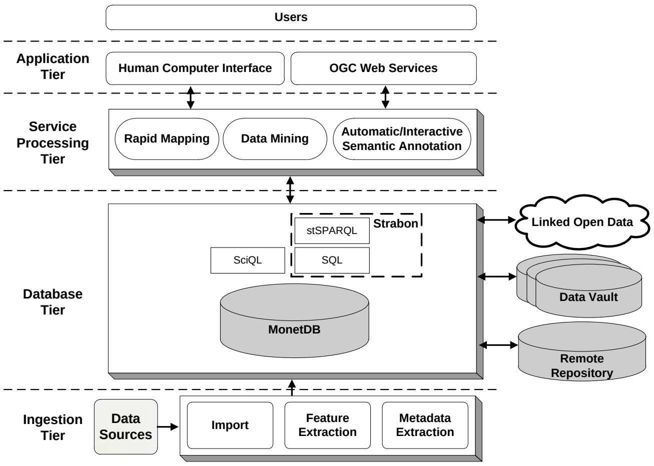


Figure 2: Software Architecture of the TELEIOS Virtual Earth Observatory

tic 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.

3. SYSTEM ARCHITECTURE

In this section we present the software architecture of the Virtual Earth Observatory. A high level view of the software architecture is presented in Figure 2. The components presented in Figure 2 can be categorized into four tiers according to the functionality that they provide:

(1) Ingestion Tier: It consists of components that perform data ingestion and content extraction.

The data ingestion components transform the original satellite image into a representation as table or array that the DBMS can handle. As a result, the optimization and execution engines of the DBMS gain transparent access to the image content instead of treating it as a "black box" BLOB or a file external to the database. In addition, the ingestion components perform operations like cropping an image to keep only the area of interest and georeferencing an image to a specific coordinate reference system.

The content extraction components perform feature and metadata extraction from raw files and products. The components that perform feature extraction create a set of patches by cutting images into square patches. In addition, they may also create feature vectors, implying that data shall be compressed into a compact multi-element feature vector representation. The components that extract metadata from raw data and products describe them using hierarchies of domain concepts that are formalized using OWL ontologies.

(2) Database Tier: It consists of components that provide access to data, metadata and semantic annotations. The main components in the DBMS tier are the systems MonetDB and Strabon. Data and metadata that were extracted from the input data during the ingestion phase are stored in MonetDB. Data of various processing levels are stored in MonetDB as multi-dimensional arrays. Transparent access to data and metadata is made possible with the use of SciQL. SciQL provides efficient array manipulation primitives that enables us to perform low level image processing and image content analysis using a high-level declarative query language.

Metadata and semantic annotations are stored in Strabon which utilizes MonetDB as a relational backend (or a spatially enabled DBMS such as PostGIS). Metadata and semantic annotations are expressed in RDF 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 using stSPARQL.

A more generic solution that addresses the principal problem of ingestion of data from external file formats into database tables or arrays, called the Data Vault [6], is used in the context of TELEIOS. 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 database tables or arrays inside the database.

(3) Service Processing Tier: It consists of Rapid Mapping services, Data Mining services and services for Automatic or Interactive Semantic Annotation. In this demo, only the Rapid Mapping services are utilized. The Rapid Mapping services expose the functionality offered by the lower layers to the application tier. These services enable an application to execute the processing chain of NOA using SciQL, to improve the thematic accuracy of the generated products using stSPARQL, and to interactively generate a map enhanced with auxiliary linked data sources.

(4) Application Tier: It consists of applications and services that provide domain specific support to the end user community.

4. DEMONSTRATION OVERVIEW

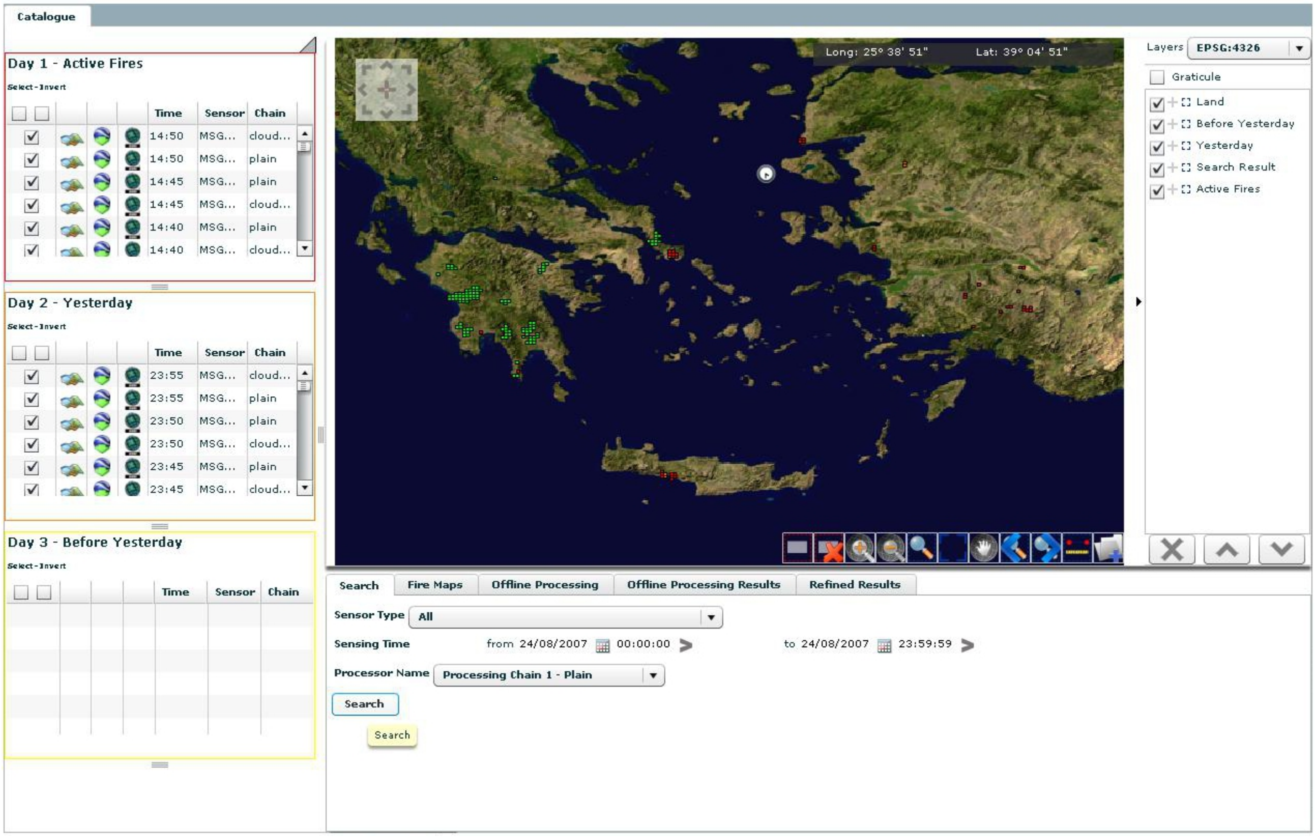
We will give a detailed example of a fire monitoring application that we have just completed using TELEIOS technologies. The National Observatory of Athens (NOA) has been archiving and processing on a routine basis large volumes of satellite images of different spectral and spatial resolutions in combination with auxiliary geo-information layers (land use/land cover data, administrative boundaries, roads and infrastructure networks data) to generate, validate and deliver fire-related products and services. In this context NOA has been developing a real-time fire hotspot detection service for effectively monitoring a fire-front.

This service consists of two categories of operations that we will demonstrate as separate scenarios.

The NOA processing chain

The processing chain utilized by the NOA fire monitoring service consists of the following modules: (a) ingestion, (b) cropping, (c) georeference, (d) classification, and (e) generation of shapefiles containing the geometries of hotspots.

In the context of this scenario the user will navigate through the GUI of the Virtual Earth Observatory depicted in Figure 3 to launch an instance of the processing chain, using a subset of the available raw data as input. She will also use the search facilities of the Virtual Earth Observatory to retrieve raw data and derived products from previous executions of a processing chain. Thus, she will test the efficiency of different processing chains (i.e., chains using a different classification submodule) by comparing products generated



Day 1 - Active Fires

Select - Invert

□ □				Time	Sensor	Chain
☑				14:50	MSG...	cloud...
☑				14:50	MSG...	plain
☑				14:45	MSG...	plain
☑				14:45	MSG...	cloud...
☑				14:40	MSG...	plain
☑				14:40	MSG...	cloud...

Day 2 - Yesterday

Select - Invert

□ □				Time	Sensor	Chain
☑				23:55	MSG...	cloud...
☑				23:55	MSG...	plain
☑				23:50	MSG...	plain
☑				23:50	MSG...	cloud...
☑				23:45	MSG...	plain
☑				23:45	MSG...	cloud...

Day 3 - Before Yesterday

Select - Invert

□ □				Time	Sensor	Chain

Long: 25° 38' 51" Lat: 39° 04' 51"

Layers EPSG:4326

□ Graticule
☑ Land
☑ Before Yesterday
☑ Yesterday
☑ Search Result
☑ Active Fires

Sensor Type: All

Sensing Time: from 24/08/2007 00:00:00 to 24/08/2007 23:59:59

Processor Name: Processing Chain 1 - Plain

Figure 3: TELEIOS Virtual Earth Observatory GUI

by each one of them.

We will also demonstrate to the users how SciQL queries are used to implement the NOA processing chains, and how stSPARQL queries are used for the discovery of raw data and derived products.

Improving generated products

An important issue in NOA's fire monitoring application is the improvement of the thematic accuracy of the outputs (shapefiles) of the hotspot processing chain. In TELEIOS, the thematic accuracy of these shapefiles is improved automatically with an additional post processing step that refines them transforming them into RDF and comparing them with relevant geospatial data also available in RDF.

Through this refinement step we isolate parts of the geometries of the hotspots that are inconsistent with the geospatial data available, but have been classified as hotspots earlier due to the low spatial resolution of the MSG/SEVIRI sensor used. During this scenario, the user will be presented with the stSPARQL update statements used to execute this refinement process and will be able to observe the effect of each step of the process visually.

Finally, we will demonstrate how the automatic generation of fire maps enriched with relevant geo-information available as open linked data is made possible with the use of a series of stSPARQL queries and the visualization of the results. This automatic generation is of paramount importance to NOA, since the creation of such maps in the past has been a time-consuming manual process.

Acknowledgments

This work has been funded by the FP7 project TELEIOS (257662), partly in cooperation with the FP7 network-of-excellence PlanetData (257641).

5. REFERENCES

[1] Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: ISWC. pp. 731–746 (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] Dumitru, C.O., et .al: KDD concepts and methods proposal: report & design recommendations. Del. 3.1, TELEIOS project (2011)

[5] Garbis, G., et al.: An implementation of a temporal and spatial extension of RDF and SPARQL on top of MonetDB - phase I. Del. 4.1, TELEIOS project (2012)

[6] Ivanova, M., Kersten, M.L., Manegold, S.: Data Vaults: a Symbiosis between Database Technology and Scientic File Repositories. In: SSDBM (2012), (Accepted for publication)

[7] Koubarakis, M., Kyzirakos, K.: Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL. In: ESWC (2010)

[8] OGC: GeoSPARQL - A geographic query language for RDF data (November 2010)

[9] Zhang, Y., Kersten, M.L., Ivanova, M., Nes, N.: SciQL: bridging the gap between science and relational DBMS. In: IDEAS. pp. 124–133 (2011)