THE MELODIES PROJECT: INTEGRATING DIVERSE DATA USING LINKED DATA AND CLOUD COMPUTING

Jon Blower[1]*, Debbie Clifford*[1]*, Pedro Gonçalves*[2] *and Manolis Koubarakis*[3]

[1] Department of Meteorology, University of Reading, United Kingdom
[2] Terradue Srl, Roma, Italy
[3] Dept. of Informatics & Telecommunications, National & Kapodistrian University of Athens, Greece

ABSTRACT

We present an overview of the MELODIES project, which is developing new data-intensive environmental services based on data from Earth Observation satellites, government databases, national and European agencies and more. We focus here on the capabilities and benefits of the project's "technical platform", which applies cloud computing and Linked Data technologies to enable the development of these services, providing flexibility and scalability.

Index Terms**—** linked data, semantic web, open data, data analytics, cloud computing, data and information visualization

1. INTRODUCTION

Environmental applications of space-derived data typically require the integration and fusion of multiple diverse datasets. Increasingly many such datasets are being made available at no cost and with liberal licences enabling wide reuse. These data encompass both scientific data about the environment (from the ESA Climate Change Initiative, the present and future Sentinel missions and from other observing systems) and other public sector information, including diverse topics such as demographics, health and crime. Many open geospatial datasets (e.g. land use and mapping) are already available through the INSPIRE directive and made available through infrastructures such as the Global Earth Observation System of Systems (GEOSS). The potential value inherent in open data, and the benefits that can be gained by combining previously-disparate sources of information, are only just starting to become understood. The MELODIES project (*Maximizing the Exploitation of Linked Open Data In Enterprise and Science*) is providing impetus in this field, stimulating the use of open data in real-world scenarios.

This presentation will give an overview of the innovations of the MELODIES project, in which we focus particularly on two particular aspects of Big Data: the handling of large data *volumes* in a scalable fashion using cloud computing, and the handling of data *variety* using Linked Data techniques.

The phrase "Linked Data" describes a set of best-practice approaches for publishing data on the World Wide Web. Linked Data are published in machine-readable form using open standards from the World Wide Web Consortium (W3C), usually under open licences. A key feature is that datasets are linked together using unique references (*identifiers*), enabling users to easily discover information that is related to their field of investigation. Many governments world-wide are beginning to publish public data in these forms to enable innovative value-adding services to emerge.

We apply these techniques to the development of eight innovative and sustainable environmental services in a broad range of societal benefit areas. The services are diverse but follow a similar general approach: the processing of Earth Observation data using distributed computing to extract information, the integration of these outputs with other data sources (e.g. government and mapping data), and the visualization of the combined results in a web portal (see figure 1). Anticipated users of these services include government and the public sector (e.g. monitoring greenhouse gas emissions or ocean water quality) and the private sector (e.g. precision farming and shipping). A key feature of these services is that they are not technology demonstrators, but are intended to be future operational services. The development of most of these services is led by SMEs. (The MELODIES consortium consists of sixteen partners from eight countries; nine of these partners are SMEs).

This paper provides a brief overview of the MELODIES "technical platform", which is a shared infrastructure that enables the development of these services.

2. OVERVIEW OF TECHNICAL PLATFORM

The technical platform includes software for managing and visualizing many different kinds of data, built on top of a cloud computing infrastructure. This infrastructure has been developed over a number of years, starting with the ESA Grid Processing on Demand (G-POD) project and evolving further

This research has received funding from the European Union Seventh Framework Programme (FP7/2013-2016) under grant agreement number 603525.

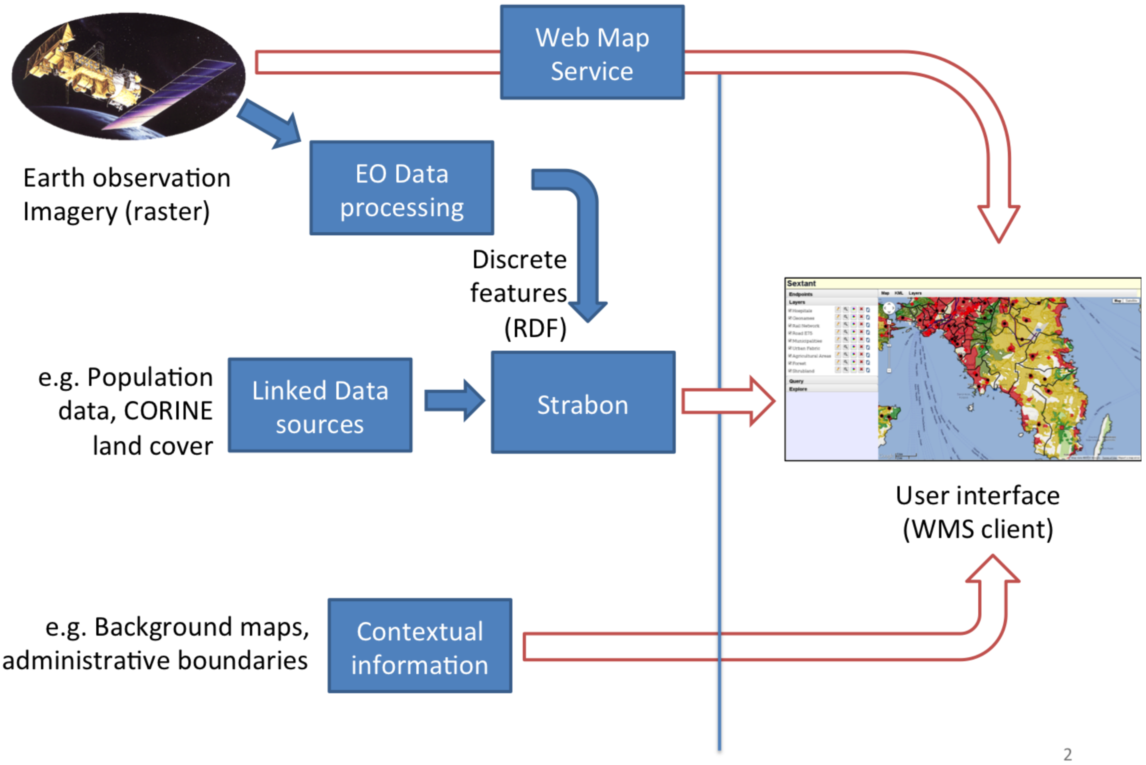
Web Map Service
Earth observation Imagery (raster)
EO Data processing
Discrete features (RDF)
e.g. Population data, CORINE land cover
Linked Data sources
Strabon
Sextant
User interface (WMS client)
e.g. Background maps, administrative boundaries
Contextual information
2

Fig. 1. The structure of a typical MELODIES service. Diverse data sources are brought into a common data store and queried in a highly flexible manner using the SPARQL query language. The results of queries are shown in a Web-GIS interface (SexTant). Large raster datasets cannot be stored in the Strabon data store, but can be served to the web interface using Web Map Services. Alternatively, discrete features can be extracted from the raster datasets and stored in Strabon.

in projects such as GENESI-DR, GENESI-DEC, GeoWOW and SenSyf. It provides the following main advantages to service developers:

1. the ability to prototype applications in a secure "sandbox" environment, allowing easy scale-out onto production systems when ready;
2. the ability to distribute data-intensive computing tasks over a cluster, using the Hadoop framework (e.g. [1]);
3. the ability to share selected datasets and components among the MELODIES services, reducing duplication of effort; and
4. the ability to control costs by only provisioning the resources that are needed at any given time, freeing unused resources for use by others.

MELODIES is also using and developing *Strabon* [2], a unique open-source system for storing and analysing time-evolving geospatial Linked Data. Strabon[1] has been used successfully for applications including burnt scar mapping and fire detection in previous European projects including TELEIOS. Strabon provides the facility to integrate datasets and query them simultaneously, providing an innovative means to generate new information and knowledge from existing data.

The MELODIES platform will also provide tools and libraries to enable quick visualization and exploration of a variety of data sources. The *SexTant* tool [3] provides a web-based "mash-up" system for linked geospatial data, enabling rapid application development and prototyping. It allows the creation of thematic maps by combining linked geospatial data and other geospatial information available in vector or raster formats (e.g., KML, GeoJSON, GeoTIFF). The MELODIES platform also provides Web Map Services (including ncWMS [4]), which enable visualization of large raster data volumes, such as satellite and numerical model data). Figure 1 gives an architectural overview of a typical MELODIES service.

3. KEY INNOVATIONS

The MELODIES project is addressing a number of current research challenges in the handling of large Earth Observation datasets. Examples of these capabilities will be demonstrated in the presentation that accompanies this paper.

3.1. Improving processing chains

The technical platform improves the development and execution of complex EO processing chains in two main ways:

1. The use of distributed computing and exploitation of the elastic nature of cloud computing means that data processing tasks can easily be scaled up to use more computing resources, either to process larger datasets or reduce the total processing time. Services can be developed to serve a local or national audience in the knowledge that they can be scaled up to continental or global audiences if required, opening up new markets.
2. The use of Linked Data techniques and the Strabon/SexTant system enables service developers to rapidly experiment with new combinations of data without the need to write new code each time: instead, they can simply formulate a new query (in the SPARQL[2] query language or a derivative language) to the data store. This means that processing chains do not need to be completely fixed in advance, and can be modified "on the fly" during the experimental phase of development.

3.2. Discovery of new knowledge from diverse datasets

The use of Linked Data also enables the usability of the large volume of the available geospatial and Earth Observation data. Semantic Web technologies allow getting value from the data by posing sophisticated queries that combine previously disparate data sources and enable new knowledge to be revealed, without the need to write new programs each time. Analysts are able to pose complex queries in the SPARQL query language and visualize the results or create diagrams that are useful for data analytics. The technical details of the original datasets are largely hidden from the analyst, allowing him/her to focus on the problem at hand. In this way, an even larger group of end user applications are able to efficiently exploit open data.

3.3. Advanced data searching across catalogues

MELODIES is driving innovations in the field of "correlation searches", which integrate several search result feeds coming from different search engines, creating relations between the different entries. This will aim ultimately at building "data casting" applications to drill down into distributed catalogues and define a coherent set of data of interest for a user of a MELODIES Service. For example, results from a flight track search engine containing the co-location parameters allow performing subsequent spatial queries to an Earth Observation search engine to discover the co-location pairs of satellite imagery to a given flight track. Therefore, sophisticated and flexible queries for precise data discovery become possible.

The core capability of correlation searches is the definition of operations that analyse not only the spatial and temporal footprints of the resources but also other data properties.

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

[2] http://en.wikipedia.org/wiki/SPARQL

Correlation search helps scientists to better define their data discovery workflow and to drill down into distributed catalogues defining a coherent set of data of interest for an experiment. The goal is to no longer focus on the metadata discovery or data download but on service providing data casting from independent, multiple sources and data staging from storages towards Cloud processing nodes.

We are starting this work from previous EC- and ESA-funded activities on extensions to the OpenSearch standard. In this way (and others), the project is contributing to the highly active area of enabling interoperability between Linked Data standards (from the W3C), geospatial standards (from bodies such as the Open Geospatial Consortium, OGC), and Internet standards from the Internet Engineering Task Force (IETF). Such interoperability has the potential to allow application developers to break out of domain-specific "silos" and harmonize data from very diverse sources.

4. KEY CHALLENGES

A number of key challenges remain to the project. One major technical challenge is the handling of large gridded data volumes (e.g. Earth Observation or numerical model data) in a Linked Data environment. Such datasets are unsuitable for conversion to RDF format as the data volumes would increase hugely. Nevertheless, the ability to query gridded data simultaneously with non-gridded data in stSPARQL queries would be extremely useful and productive, enabling the easy combination of models, satellite data and in situ observations. We are experimenting with a number of techniques, including treating the gridded dataset as a "virtual" RDF store, then implementing a SPARQL query engine above it. Another strategy is to extract discrete features from the EO data before storing in the Strabon data store (as in figure 1).

The challenge of ensuring the long-term *sustainability* of the MELODIES services and technologies is a very strong one. The project includes a workpackage that is systematically investigating the factors affecting sustainability, including economic, political and technical trends. The goal is to consider issues of sustainability from the earliest stages of development.

5. CONCLUSIONS

The MELODIES project is stimulating the use of Open Data, including Earth Observation data, by developing eight real-world applications, using a shared technology platform to enable research groups and small and medium-sized business to build innovative data-intensive services that address a variety of societal needs. Through the use of cloud computing and Linked Data techniques, we can better handle problems of data volume and diversity, and increase the flexibility of our applications.

6. REFERENCES

[1] Francesco Casu, Stefano Elefante, Pasquale Imperatore, Riccardo Lanari, Michele Manunta, Ivana Zinno, Emmanuel Mathot, Fabrice Brito, Jordi Farres, and Wolfgang Lengert, "DInSAR time series generation within a cloud computing environment: from ERS to Sentinel-1 scenario," in *European Geosciences Union conference*, 2013, EGU2013-8365.

[2] Kostis Kyzirakos, Manos Karpathiotakis, and Manolis Koubarakis, "Strabon: A semantic geospatial DBMS," in *International Semantic Web Conference (1)*, 2012, pp. 295–311.

[3] Konstantina Bereta, Charalampos Nikolaou, Manos Karpathiotakis, Kostis Kyzirakos, and Manolis Koubarakis, "SexTant: Visualizing time-evolving linked geospatial data," in *Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*, 2013, pp. 177–180.

[4] J.D. Blower, A.L. Gemmell, G.H. Griffiths, K. Haines, A. Santokhee, and X. Yang, "A web map service implementation for the visualization of multidimensional gridded environmental data," *Environmental Modelling & Software*, vol. 47, pp. 218–224, Sept. 2013.