THE ESPAS E-INFRASTRUCTURE: ACCESS TO DATA FROM NEAR-EARTH SPACE

Anna Belehaki (1), Mike Hapgood (2), Natalia Manola (3), Spiros Ventouras (2), Sarah James (2), George Athanasopoulos (3), Antonis Lempesis (3), Stefania Marziou (3), Jurgen Watermann (4), and the ESPAS team (5)

(1) University of Athens and National Observatory of Athens - http://www.iono.noa.gr
(2) RAL-STFC - http://www.espas-fp7.eu/index.php/community/partners/7-stfc
(3)ATHENA Research Center - www.athena-innovation.gr
(4) jfwConsult - http://jfwconsult.free.fr/
(5) http://www.espas-fp7.eu/trac/wiki/PublicPages/ESPASConsortium

ABSTRACT

The aim of the ESPAS e-Infrastructure is to facilitate access to data from the near-Earth space environment enabling exploitation of multi-instrument, multi-point science data for analysis, model building, data assimilation into models, model-observation comparison, space environment nowcast and forecast, to name just a few.

The ESPAS project, currently in its fourth year of activities, has developed a set of tools to allow the data providers to interoperate through well established standards, deployed in at all participating data nodes and a coordinating central platform. This paper provides a high-level overview of the system model, architecture and functionalities, and presents indicative science use cases that ESPAS can support.

Index Terms— data infrastructure, near-Earth space, interoperability, OGC - Open Geospatial Consortium

1. INTRODUCTION

The “Near-Earth Space Data Infrastructure for e-Science" (ESPAS) develops a platform to facilitate access to near-Earth space data. These are heterogeneous data from a variety of different sensors (several sensors on the CLUSTER, DEMETER, NOAA-POESS, ACE, MAGION3 spacecraft and on ground, including ionosondes, magnetometers, Fabry–Pérot *interferometers*, Incoherent Scatter Radars, SUPERDARN) and model output data (e.g., EDAM, CMAT, IRI), supplied by a number of widely distributed institutions with different data acquisition and processing philosophies and different data distribution policies. From its inception, ESPAS was designed to meet the following requirements:

- Integrate heterogeneous data from multiple providers, ranging from ground-based observations acquired with multiple instruments to data from spaceborne sensors. To do so, specific policies are established on identification, access, availability, quality, sharing and re-use of the data (or metadata) of the participating content providers.
- Enable data search and access across multiple data sources through a central platform. This is possible through the establishment of a variety of workflows (data flows) initiated on the ESPAS platform.
- Provide value-added services to explore metadata, visualize or manipulate the integrated data, and eventually mine metadata (events identification).
- Serve as a test-bed for proposed methodologies and standards for model validation and optimization. This will lead to validated predictions of the conditions in the near-geospace environment, able to support operational applications.
- Allow for extensive testing through several test and use cases, designed to serve the needs of the wide and interdisciplinary user and provider communities for computationally intense science carried out in a highly distributed data environment.
- Integrate into the wider European scientific infrastructures. To do so, it carries on and extends the policies and procedures issued at its participating data providers on data openness, quality, identification, etc.

2. THE ESPAS ARCHITECTURE

ESPAS uses and adapts well-established techniques from related data e-Infrastructures disciplines as well as the Digital Libraries domain to provide the required semantic integration of the participating data sources. Main elements towards building interoperability are: the data model, which is compatible with and encapsulates all underlying data formats; the definition and employment of XML schemas for metadata exchange format; the domain-specific vocabularies used to describe the near-earth science data and phenomena; the services (*wrappers)* installed at the data nodes to support metadata publishing and data exposition; the central coordinating platform that showcases the metadata aggregation and data access through a portal that provides tools for data registration and validation and customized search workflows for more complex queries. The current ESPAS release provides access to the majority of datasets registered in ESPAS from space and ground based assets and from scientific models.

Based on similar approaches applied in other scientific data infrastructures, the ESPAS platform architecture is a

doi: 10.2788/1823

modular, Service Oriented Architecture (SOA), based on the D-NET system [3]. It is comprised of the following logical layers as illustrated in **Figure 1**:

– *Enabling Layer:* provides the necessary components that glue, manage and organize all the services in an SOA. It implements service registration and a notification based communication system, making the overall system extensible and extendible.
– *Data Management Layer:* focuses on the metadata harvesting and the basic data management services of the ESPAS system: storage, database, index.
– *Access and Value-added Services Layers:* Include advanced search, statistics capabilities, and enable the development of models and visualizations tools.
– *Web Layer:* provides the User Interface service, i.e., the ESPAS portal.

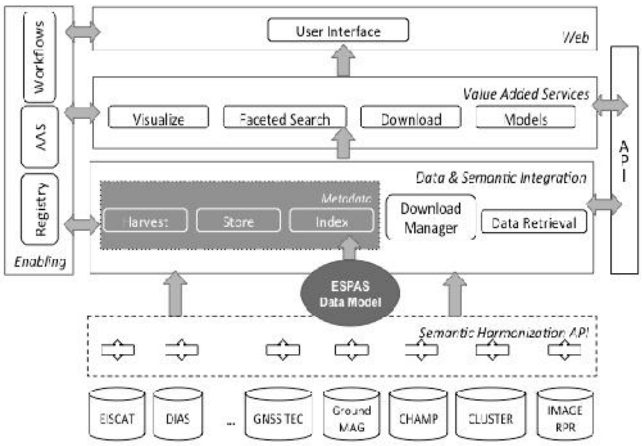

Figure 1. ESPAS overall architecture.

The *Semantic Harmonization Layer* binds the ESPAS platform with the data providers and implements the harmonization of the participating data sources through a set of services that adopt the ESPAS data model on well established formats and transfer protocols (XML, OGC).

2.1 ESPAS data model

A major step for achieving interoperability among the different data formats is the development of a common data model able to handle all heterogeneous data files to be accessed by ESPAS[1]. Using this model, data providers are able to fully describe and export metadata relevant to their measurements and provide semantic information about the measurements themselves. For the development of the ESPAS model we have adopted the concept of *feature* (ISO 19101:2002; ISO 19109:2005 for Geographic Information), which represents an abstraction of real-world objects. Consequently a portion of the near-earth world is modeled in terms of features together with their properties.

The ESPAS model is comprised of two parts. Part 1 focuses on the **observation** (ISO 19156:2011 and OGC Observations and measurements[2]), an entity associated with a discrete time instant or period through which a number, term or other symbol is assigned to a **phenomenon**. It indicates the descriptive metadata such as its provenance (e.g., instrument, model, program, organization), quality (e.g., instrument calibration, originating process) and grouping denoting logical datasets (e.g., experiments, time-series data). Part 2 focuses on the result of an observation, i.e., an **observed property,** which is an estimate of the value of a property of some feature attached to the observation

The observed property and its characteristics are an entry point to the description of the Near-Earth space domain. The model describes the relationships between concepts (namely, phenomena and observed properties) in the Near-Earth space and how these possibly relate to similar domains. ESPAS is using controlled vocabularies throughout for the different terms and definitions (e.g., instrument types, platforms, roles, phenomena, units) currently converted into SKOS format (http://www.w3.org/2004/02/skos/), a standard for dictionaries/thesauri that also supports hierarchical relationships.

2.2 ESPAS Interoperability services

To facilitate access to data offered by data providers a set of basic supporting services has been defined and implemented by ESPAS. Each of these services essentially implements a client-server protocol, with the server part residing at the data provider side (in the *Semantic Harmonization Layer,* bundled within the ESPAS wrapper) and the client side at the ESPAS platform *Data Management Layer*:

- An **OGC compliant Catalogue Service (CSW),** which supports the identification of ESPAS resources offered by each data provider.
- A **Download Service**, that facilitates the download of data bundles in terms of data collections offered by each provider.
- An **OGC Compliant Sensor Observation Service (SOS)** which facilitates the collection of selected data parameters or values from the observations of each data provider.

HARVESTING METADATA: OGC CATALOGUE SERVICE

OGC Web services (OWS) are the prevailing types of services in the geospatial application domain[4]. Based on the OGC Catalogue Service for the Web (CSW)[5], i.e., a profile of the catalogue service with the goal to expose the catalogue functionality over the web, ESPAS has partially implemented the specified CSW interface specification to accommodate the `getRecords` operation. Through the `getRecords` operation, clients are able to submit queries for the discovery of metadata records for each of the major metadata entities specified by the ESPAS model. Constraints on the expected results are expressed as Contextual Query Language - CQL queries over a list of specific properties, which include, the type and the modification date of the expected metadata entity. The supported binding includes HTTP Key Value Pair (KVPs).

RETRIEVING DATASETS: DOWNLOAD SERVICE

The download service is an ESPAS proprietary service that enables ESPAS clients to retrieve data collections in formats defined by the data providers and specified in the metadata descriptions. It is implemented on top of basic D-NET services, which provide subscribe-notify and database interaction operations, as well as on other services that facilitate policy enforcement functionality. The download service operates in an asynchronous manner so as to avoid blocking clients from performing other activities while the download request is submitted and processed by the associated data providers. In terms of architecture[6], the download service may be logically split into three distinct components, namely the Central Download Manager, local Download Service and Policy Manager. The Central Download Manager component is responsible for checking conformance of each request against the policies defined in the Policy Manager and coordinating the interaction with the local Download Services. The Policy Manager is an XACML v3.0 [6] based policy server, responsible for maintaining the policies defined by the data providers and validating each request against related policies. Local Download Services are deployed on each provider and are responsible for processing the download requests related to each distinct provider.

RETRIEVING DATA VALUES: OGC SOS SERVICE

One of the well-known types of OGC Web services is the Sensor Observation Service (SOS). SOS is a web service that is used to query real-time sensor data and sensor data time series and is part of the Sensor Web vision. The offered sensor data comprises descriptions of sensors themselves, which are encoded in the Sensor Model Language (SensorML) and the measured values in the Observations and Measurements (O&M)[7] encoding format.

The preliminary implementation offered by ESPAS is supporting partially the specified SOS service interface[8]. It includes a customized implementation of the `getResult` and `getResultTemplate` operations, which facilitate the retrieval of specific observed property values and value templates respectively, for observations performed by data providers. Among the set of optional query criteria defined for the `getResult` operation the provided implementation supports only time related ones, whereas in terms of bindings only HTTP Key Value Pair (KVPs) based requests.

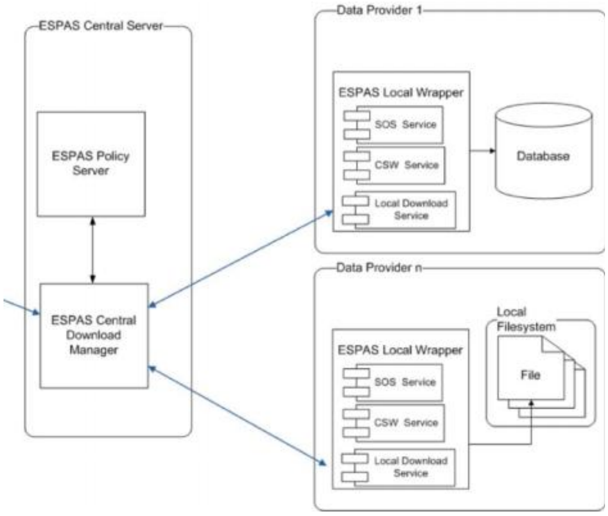

Figure 2. Download Manager

Moreover, the actual implementation requires a layer that retrieves data from the data providers, whether this might reside in databases or files. Using the HELIO approach, and bundled in the ESPAS wrapper, the ESPAS implementation of the SOS service is able to do one of the following through the use of generic services that rely on configuration: convert standard SQL queries, HELIO-compliant databases, or retrieve from files through the CSX generic services.

3. THE ESPAS PORTAL

The ESPAS User Interface is accessible from the public web site http://www.espas-fp7.eu and offers three major functionalities:

(1) **data registration and validation**: Data providers use the service to describe their datasets according to the ESPAS model, produce the corresponding XML and upload it to the ESPAS catalogue service. Validation is enabled throughout the registration process, assisting the providers in understanding the concepts and preventing them from introducing errors into the final catalogue;

(1) **faceted and progressive search:** End users search the ESPAS catalogue using *Time*, *Assets*, *Observed Properties* and *Observation Collections* to progressively filter their search *in real time*. Location based queries are currently computed off-line, since the combination of observation locations from ground stations and satellites becomes a complex issue involving different location coordinate systems which involve time, conjunctive statements, all in an information space that may have time gaps and which requires further optimizations and

approximations. All results are provided through observations;

(3) **datasets and data values download request**: Users are able to select and group datasets or values of observed properties and request their download. This process is carried out by the Download Manager Service, which, after checking against all relevant data provider policies, launches requests to the corresponding data nodes. For datasets, each provider bundles all files and passes on a URI for their retrieval. For data values, the provider is returning the requested values via the OGC-SOS service, all bundled together at the central platform.

4. INDICATIVE USE CASES

Every new system needs to be tested before opening it to the external users. Therefore “Use Cases” based on specific scenarios have been defined and analyzed by the ESPAS consortium. The results will guide us to define final value-added services to be provided by the ESPAS platform.

In this contribution three indicative use cases are presented, the first is a simple functionality test, the two other address problems in research and application.

Use case A: *Search for data products*

This Use Case demonstrates the search functionality. It is essentially a data product search where ESPAS returns metadata information and links to data products at the data provider sites. Three steps are shown: (1) select a specific "observed property" from among those offered by ESPAS (in this case "electron density"), (2) search for "assets" available in ESPAS which can deliver this specific "observed property", (3) query the system which - if the query is successful - returns a list with links to "observation collections" which contain the requested observed property.

Use Case B: *Real-time mapping of ionospheric characteristics*

The purpose of this use case is twofold: (a) use ESPAS ionospheric sounder data from European high latitudes to expand the maps of foF2 and electron density currently generated in DIAS, (b) assess the potential of ESPAS to provide these maps as a value-added service in real time. These maps are hybrid products, they are constructed from scaled bottomside ionograms with model extension into the topside ionosphere.

Use case C: *Validation of a model data assimilation system*

A specific atmosphere model (CMAT2) will run several scenarios each covering several weeks of middle and upper neutral atmosphere simulation, supported by a data assimilation system. When data types available in ESPAS match those used by the CMAT2 assimilation system the data provided by ESPAS will serve as independent validation data for the model.

5. CHALLENGES TOWARDS THE 4 Vs of BIG DATA

ESPAS manages over 30 years of observations acquired with a large number of heterogeneous instruments or generated via numerical models (*variety*). The ESPAS model is able to homogenize them, but the *volume* of the observations required an upscaling to accommodate millions records, each one describing varying time periods and measuring a different region of space, often with gaps or errors in the information (*veracity*). The challenge was to store the catalogue and provide *accurate* and *fast* access to it (*velocity*) and also to implement a data value retrieval service that can potentially transfer (stream) and integrate millions of rows from a large number of data nodes. Extra complexity was added due to the fact that spatial information is described in different coordinate systems (depending on the specific scientific domain) with ESPAS validating and homogenizing it.

Acknowledgements: ESPAS is funded through EU-FP7 under grant agreement 283676.

6. REFERENCES

[1] Ventouras, S., "Common Data Model for ESPAS (version 2)", Available at http://www.espas-fp7.eu/index.php/overview/deliverables, 2013

[2] ISO 19156:20111 Observations and measurements, http://www.iso.org/iso/catalogue_detail.htm?csnumber=32574, OGC Observations and measurements Schema, http://www.opengeospatial.org/standards/om

[3] Manghi P. et al., (2014) "The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures", Program: electronic library and information systems, Vol. 48 Issue: 4, pp.322 – 354

[4] Doyle A. et al., "Introduction to OGC Web Services," OGC- Open Geospatial Consortium, OGC White Paper 2001.

[5] Nebert D. et al., OpenGIS® Catalogue Services Specification v2.0.2, Open Geospatial Consortium Standard, OGC 07-006r1, 2007

[6] Rissanen, E. (2013). eXtensible Access Control Markup Language (XACML) v3.0. OASIS standard, Available at: http://docs.oasis-open.org/ xacml/3.0/xacml-3.0-core-spec-en.html.

[7] Cox, Simon, "Observations and measurements." Open Geospatial Consortium Best Practices Document. OGC, 2006.

[8] Bröring A. et al., OGC Sensor Observation Service Interface Standard v2.0.0, Open Geospatial Consortium Standard, OGC 12-006, 2012