On Harmonizing Earth Science Policies, Semantics, Metadata and Ontologies

Yannis Marketakis[(1)], Yannis Tzitzikas[(1)], Calogera Tona[(2)], Massimo Argenti[(2)], Fulvio Marelli[(2)], Mirko Albani[(2),] Raffaele Guarino[(3)], Barbara Polsinelli[(3)], Roberto Bitto[(3)]

*[(1)] **Institute of Computer Science (FORTH-ICS)***
N. Plastira 100, Vassilika Vouton, GR-711 10 Heraklion, Crete, Greece
EMail: marketak@ics.forth.gr, tzitzik@ics.forth.gr

*[(2)] **European Space Agency - Centre of Earth Observation (ESA-ESRIN)***
Via Galileo Galilei, 00044 Frascati, Rome, Italy
EMail: calogera.tona@esa.int, massimo.argenti@esa.int, fulvio.marelli@esa.int, mirko.albani@esa.int

*[(3)] **CapGemini Italia SPA***
Via di Torre Spaccata 140, 00173, Rome, Italy
EMail: raffaele.guarino@capgemini.com, barbara.polsinelli@capgemini.com, roberto.bitto@capgemini.com

ABSTRACT

The volume of scientific research results in Earth Science is growing tremendously, so their preservation is becoming more and more appealing. Large amount of data is already affecting many fields of science, most notably fields like, space with both new satellite surveys and new deployments of extensive sensor networks, oceanography with deployments of underwater oceanographic observatories, geophysics with past and new seismograph data, etc. This trend will not be confined to the physical sciences but will also transform large parts of the humanities and social sciences.

Mechanisms, infrastructures and software solutions should be in place to enable sustainable long-term preservation of scientific results in digital form. Common preservation policies and their application by Earth Science data owners and providers should be defined to guarantee preservation of data and associated knowledge according to a common and harmonized approach and their accessibility by users according to their needs. Furthermore Earth Science communities use various models (conceptual models, metadata schemas, ontologies, vocabularies) for describing their datasets. Ideally we would like a common set of models for Earth science data, and a common strategy for harmonizing them.

The purpose of this paper, which summarizes some of the work done in the context of the SCIDIP-ES project, is to describe the Earth Science needs, procedures and gaps in terms of the current data preservation and data access policies, perform an analysis of the current Earth Science data infrastructures, identify the gaps with respect to the Earth Science community requirements and interoperability aspects, and define common data preservation policies which are applicable to all Earth Science data categories. Furthermore an analysis of the semantics, metadata and ontologies that are currently in use by earth scientists is presented. Based on this analysis, possible strategies for having harmonized metadata, semantics and ontologies are proposed, able to satisfy the earth scientists' needs coping with different Earth Science domain approaches.

Keywords: Data Preservation, Harmonization, Policies, Metadata, Semantic Web, Interoperability.

INTRODUCTION

The volume of scientific research results in Earth Science is growing tremendously, so their preservation is becoming more and more appealing. Large amount of data is already affecting many fields of science, most notably fields like, space with both new satellite surveys and new deployments of extensive sensor networks, oceanography with deployments of underwater oceanographic observatories, geophysics with past and new seismograph data, etc. The data tsunami, which is occurring in Earth Science, is gaining a critical importance. Past and present Earth Observation missions will have collected several tenths of Terabytes of data by the end of the missions; the ESA Sentinels[1], which are going to be launched between 2013 and 2020, are expected to collect 1TB of data per day. These numbers are relatively small if compared to the Square Kilometre Array radar[2], which will be completed by 2024 and is expected to generate the same amount of data each day as the entire Internet.

The problem here does not only refer to the "scale"; the complexity is mainly in the variety and heterogeneity of the data acquired from different satellites and sensors and which are stored in different formats, and in the need for additional information/knowledge which makes these data usable across different scientific communities. This leads into the need for data interoperability, which has increased dramatically in recent years. Scientists belonging to different communities increasingly need to access and use data with which they have little or no familiarity. This is particularly evident in Earth Science, where phenomena such as climate change require different communities of scientists to work together on a common objective, sharing both data and results. This provides a challenge for Earth Science data holders and archive owners who must ensure coherent data preservation and optimum availability and accessibility of the different Earth Science data and products.

The purpose of this paper is to describe the Earth Science needs, procedures and gaps in terms of the current data preservation and data access policies, perform an analysis of the current Earth Science data infrastructures, identify the gaps with respect to the Earth Science community requirements and interoperability aspects. Furthermore an analysis of the semantics, metadata and ontologies that are currently in use by earth scientists is presented. Based on this analysis, possible strategies for having harmonized metadata, semantics and ontologies are proposed, able to satisfy the earth scientists' needs coping with different Earth Science domain approaches.

This work has been carried out in in the context of the ongoing EU project SCIDIP-ES[3] whose objective is to deliver generic services for science data preservation as part of the data infrastructure for e-science and set up a European framework for the long term preservation of Earth Science data through the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of generic infrastructure services in the Earth Science domain. This paper summarizes the results reported in the SCIDIP-ES deliverables [1,2].

THE STATE OF THE ART

Data Preservation Policies

Earth Science data are the products of different missions, campaigns or experiments carried out by different organizations and as a result are complex and heterogeneous. The rationale which has been chosen to categorize these data, and therefore to conduct an in-depth analysis of the different data preservation and access policies has been based on the classification of the sensors or instrument types which are behind these missions. This has resulted into the ten categories shown in Table 1. The first five categories (C1 to C5) can be grouped and referred to as "Earth Observation Space data"; categories C6

[1] http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4
[2] http://www.skatelescope.org/
[3] http://www.scidip-es.eu

and C7 can be grouped and referred to as “Airborne and Balloon” data, while C8 and C9 categories can be grouped and referred as “In-Situ data”. The remaining category (C10 “Models and Simulations data”) refers to Earth Science experiments and simulations which are conducted in the laboratory.

#	**Category**
C1	SAR imaging missions/sensors, high and very high resolution (different radar bands)
C2	Multi-spectral imaging missions/sensors, high and very high resolution
C3	Medium resolution land and ocean monitoring missions/sensors (e.g. wide swath ocean color and surface temperature sensors, altimeter, etc.)
C4	Atmospheric chemistry missions/sensors
C5	Satellite based other scientific missions/sensors
C6	Airborne (e.g. digital cameras single/multiple, digital line scanners, radar, laser topographic/ bathymetric, etc.). Helicopter Observation Platforms (HOPs) are considered in this category
C7	Balloon (e.g. geomagnetic instruments, wind, temperature, radiation, radio propagation, particles, optical properties, chemistry, etc.)
C8	Ground (e.g. seismography, temperature, humidity, wind, pressure, radiation, radiance, pollution factors, rain, chromatography, soil property, etc.)
C9	Hydro (e.g. temperature, salinity, pollution factors, wind, pressure, water flow/flux/level, etc.). This includes data coming from buoys as well as from ships, gliders or other equipment used to capture local data
C10	Models and Simulations data

Table 1 A categorization of Earth Science data

A complex process such as data preservation in the Earth Science domain can be divided into logical steps or “themes”. The definition of these themes has been derived from an extensive analysis of the European LTDP[4] Common Guidelines. They consist of a set of technical guidelines recommended for application by Earth Observation space data owners for the preservation of their data. These LTDP guidelines do not cover programmatic and regulatory aspects (e.g. LTDP organization, availability of LTDP dedicated programs or budgets within organizations) or data policy aspects associated with the data to be preserved. These themes are shown in the following table:

Theme #	**Description**
Theme 1	Preserved data set content definition and appraisal
Theme 2	Archive operations and organization
Theme 3	Archive security
Theme 4	Data ingestion
Theme 5	Archive maintenance
Theme 6	Data access and interoperability
Theme 7	Data exploitation and re-processing
Theme 8	Data purge prevention

Table 2 Themes for Data Preservation

For each of the data categories (described in Table 1) we performed an analysis to investigate the current practices for the preservation of Earth Science data using the aforementioned themes as criteria. The analysis revealed that for “Earth Observation space data” (C1-C5) most of the organizations have in place their own practices which are aligned with the principles defined by LTDP common guidelines. There is a strong reliance on ISO standards and well known formats (e.g. PDF) are preferred for documentation.

[4] Long Term Data Preservation http://earth.esa.int/gscb/ltdp/

Regarding “Airborne and Balloon data” (C6-C7) different agencies are working to preserve their heritage resources and making them accessible to the public. Moreover the datasets in these categories seem to be very heterogeneous and raw data or the intermediate level processing data are usually not provided. Regarding In-Situ data (C8-C9) the tendency is to store the raw data, the data products and sometimes the associated documentation; however the preservation of the software tends to be less common. Moreover there are cases where the procedures for the long-term preservation of the data have not been adopted (in such cases preservation is only addressed in terms of the physical storage of the data). Finally with respect to the Models and Simulations data (C10) there seems to be a full adherence to the guidelines, although the field is very heterogeneous.

Data Access Policies

Data access policies define what data can be accessed by what users, and under what conditions and they are often accompanied by terms and conditions of use. Data access policies from publicly funded organizations are generally aiming at providing users with access to data on a non-discriminatory basis (full and open access), at minimum (or no) cost, as early as possible in the dissemination process.

The most basic level of data access is through an index or catalogue that provides information on the existence of the data without generally granting access to the data itself (i.e. federated catalogues like ENVRI[5] or GEO[6] portals). In this case users define basic search attributes such as the area of interest, type of data, date of acquisition etc. and the portal returns metadata records for the datasets, which are available without providing a direct link to the data itself. A second level of data access is provided by services which expose a representation of the data (e.g. as an image) without giving the end user access to the original dataset. The last and most complete level of data access allows the user to download the original dataset or an equivalent representation of it.

We evaluated the data access policies adopted by several organizations using the following criteria: (i) investigation of the openness of access to the data and the eventual restrictions different communities of users/owners have, and the mechanisms described by the data policy to allow them to search, discover and access the data, trying to identify commonalities or particular solutions, (ii) analysis of the general mechanism for assigning roles and privileges for the different communities, (iii) overview of the different pricing policies adopted by the data owners or distributors trying to define/discover a general price structure, (iv) identification of legal issues related to the management of intellectual property rights for ES data, (v) investigation of agreements in distribution of data by third parties and (vi) compliance with INSPIRE[7] directives. The INSPIRE directive lays down general rules for establishing a spatial information infrastructure in Europe for the purposes of establishing community environmental policies and informing activities which may have an impact on the environment.

The analysis revealed that Data access policies are generally well consolidated in the Earth Science organisations. In particular, data tend to be available free of charge, pricing policies depend on agreements with private companies in most of the cases, intellectual property rights depend on national and international agreements, distribution of third parties data depends on national and international agreements, and the majority of organisations are either INSPIRE compliant or working towards compliance. In general the landscape in the domain of data access policies for Earth Science data is rather complex due to the abundance of different policies relating to different organizations and also within single organizations for different Earth Science data and products. This landscape depend both on the nature of the organizations (private vs. public) and on the individual national legislations and international agreements which largely influence them.

[5] http://portal.genesi-dec.eu/envri/
[6] http://www.geoportal.org/web/guest/geo_home
[7] http://inspire.jrc.ec.europa.eu/

Semantics, Metadata and Ontologies currently in use

Earth Science data are the product of different missions, campaigns or experiments carried out by several organizations, and as a result are very complex and heterogeneous. Furthermore there is not a single standard schema for maintaining and exchanging these data. In the context of the SCIDIP-ES project a survey has been carried out for identifying the models which are being used by different organizations. The results of this survey yielded about 50 distinct models. As regards their format and purpose, the majority of these data are expressed in XML format and most of them are used for querying and exchanging data. The following figure shows the distribution of the models.

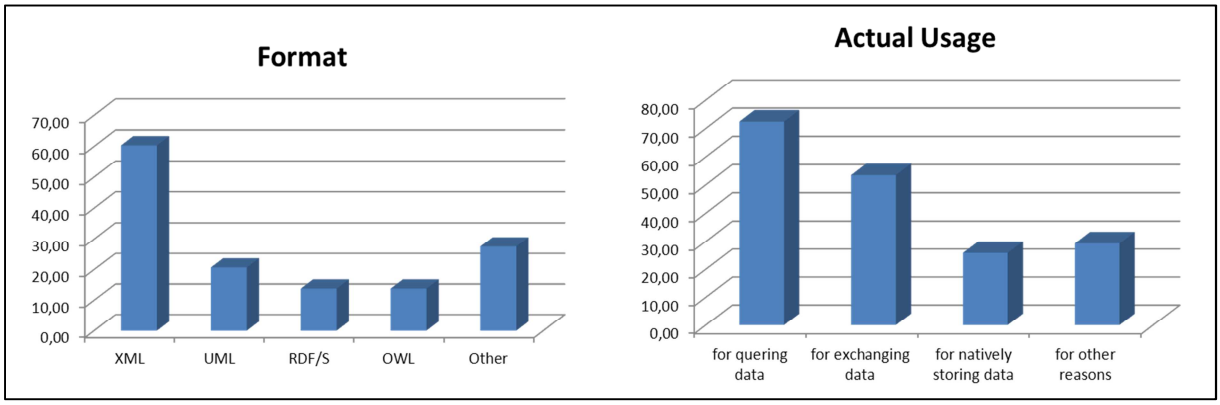

Figure 1 Format and Actual Usage Information

The next step has been to identify the most "popular" models and further analyze them. A brief description of these models is given below.

ISO 19100 Series: The International Organization for Standardization has established the ISO 19100 series[8] for: (a) defining the basic semantics and structure of geographic information for data management and data interchange purposes and (b) defining geographic information service components and their behavior for data processing purposes. In general we could say that these standards support data management, acquiring, processing, analyzing, accessing and disseminating data between different users, systems and locations for geographic information.

OGC Standards: The OGC Technical Committee has developed an architecture in support of its vision of geospatial technology and data interoperability called the OpenGIS Abstract Specification[9]. The Abstract Specification provides the conceptual foundation for most of the OGC specification development activities. Open interfaces and protocols are built and referenced against the Abstract Specification, thus enabling interoperability between different brands and different kinds of spatial processing systems. The Abstract Specification provides a reference model for the development of OpenGIS Implementation Specifications and is broken into 20 Topics in order to assist development of different topics from different working groups.

CGI Vocabularies: The development of CGI (Commission for the Management and Application of Geoscience Information) vocabularies aimed at developing concept vocabularies for populating GeoSciML interchange documents. Because the vocabularies implement GeoSciML concepts they are relevant to a number of types of geological information (e.g. Samples, Species, etc.). There are currently 31 vocabularies in the CGI portfolio.

CF Conventions: The Climate and Forecast (CF) metadata conventions[10] are designed to promote the processing and sharing of data files. In particular they are focused on the description of Earth Science

[8] http://www.isotc211.org/Outreach/Overview/Overview.htm
[9] http://www.opengeospatial.org/standards/as
[10] http://cf-pcmdi.llnl.gov/documents/cf-conventions

data. The data files are created with respect to the NetCDF API. The metadata defined by the CF conventions are included in the same file with the data, thus making the file "self-describing". CF conventions are intended for use with climate and forecast–related data.

GEMET: The General Environmental Multi-lingual Thesaurus (for short GEMET[11]) has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA). The basic idea for its development was to merge the best available multilingual thesauri. GEMET was conceived as a general thesaurus, thus specific thesauri and descriptor systems were not included in its development and only their upper level structure and terminology was taken into account.

GeoSciML: GeoSciML[12] was developed by CGI, as a Geography Markup Language (GML) derived schema to represent geology data. GeoSciML is a data interchange format which (not strictly) falls in the ontology/metadata/vocabulary domain.

OTEG: OTEG has been developed within the framework of ESA activities in the field and in line with related past and ongoing projects. The aim of the OTEG project is to help application experts in identifying relevant Earth Observation products, allowing them to easily identify the Earth Observation missions, sensors and products useful for their activity, using familiar semantic terms (i.e., terms pertaining to their application domain).

MOLES: MOLES [10] (Metadata Object for Linking Environmental Sources) is an information model for describing metadata covering a broad range of applications within multiple disciplines. These are mainly, but not limited, those within the earth and physical sciences. MOLES is primarily of use to consumers of data, especially in an interdisciplinary context. It allows them to establish details of provenance, and to compare and contrast such information without recourse to discipline-specific metadata or private communications with the original investigators.

ThIST: is an abbreviation of *"Thesaurus Italiano di Scienze della Terra"*, otherwise *"Italian Thesaurus of Earth Science"*. It is a thesaurus which is used to classify Earth Science related documentation and cartography within the internal library archives. Currently it contains about 10350 descriptors linked to each other through a set of more than 100 thousand relationships of hierarchical, associative and equivalence types.

VoID: The Vocabulary of Interlinked Datasets (VoID)[13] is a vocabulary exclusively focused on providing metadata for describing datasets as a whole. It provides terms and patterns for describing RDF datasets, and is intended as a bridge between the publishers and users of RDF data. VoID descriptions can be used in many situations, ranging from data discovery to cataloguing and archiving of datasets, but most importantly it helps users find the right data for their tasks.

SKOS: SKOS [11] is an abbreviation of Simple Knowledge Organization System, and is implemented as a series of specifications and standards to support the use of Knowledge Organisation Systems, such as thesauri, classification schemes, and taxonomies, and other similar types of controlled vocabularies, within the framework of the semantic web. As an application of RDF, SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the web and integrated into other concept systems.

Dublin Core: The Dublin Core metadata terms [12] are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. Typically, the Dublin Core vocabulary has been adopted and/or extended by many other domain vocabularies to directly refer to the domain objects represented in RDF though, as its properties can be connected to any *rdfs:Resource*, Dublin Core is also used by specific metadata vocabularies for describing RDF vocabularies and RDF Datasets, thus to provide overall information about data collections as a whole.

---

[11] http://www.eionet.europa.eu/gemet/
[12] http://www.geosciml.org/
[13] http://www.w3.org/TR/void/

ISO 21127: CIDOC Conceptual Reference Model (ISO 21127) [3] is a formal ontology of 80 classes and 132 relations describing the underlying semantics of over a hundred database schemata and structures from all museum disciplines, archives and libraries. The primary role of CIDOC CRM is to enable information exchange and integration between heterogeneous sources on cultural heritage information. It defines (and is restricted to) the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal ontology.

TOWARDS HARMONIZING METADATA, SEMANTICS AND ONTOLOGIES

The "Big Picture" conceptually

Since the existing Earth Science resources can be found in a plethora of different forms (from research papers, databases, files, catalogues, etc.), we can make the following basic distinctions: (A) products of human activities, (B) human activities and (C) core conceptualization of Earth Sciences. The above distinctions and the relationships between these models can be characterized by a core conceptual model of fundamental categories and their relationships. Figure 2 illustrates these groups formed by these distinctions, as well as the concepts they contain and the relationships between them. It is a kind of high level picture of the relationships between human activities and their targets, as they appear on the highest level of description of all scientific products and services. They are the ones necessary for the first level selecting relevant information, and for managing and maintaining the referential integrity of the most fundamental contextual information. It is only on a more specialized level that all the entities in these models are refined by specific, open-ended terminologies (typologies, taxonomies) and a relatively small set of more specific relationships among them

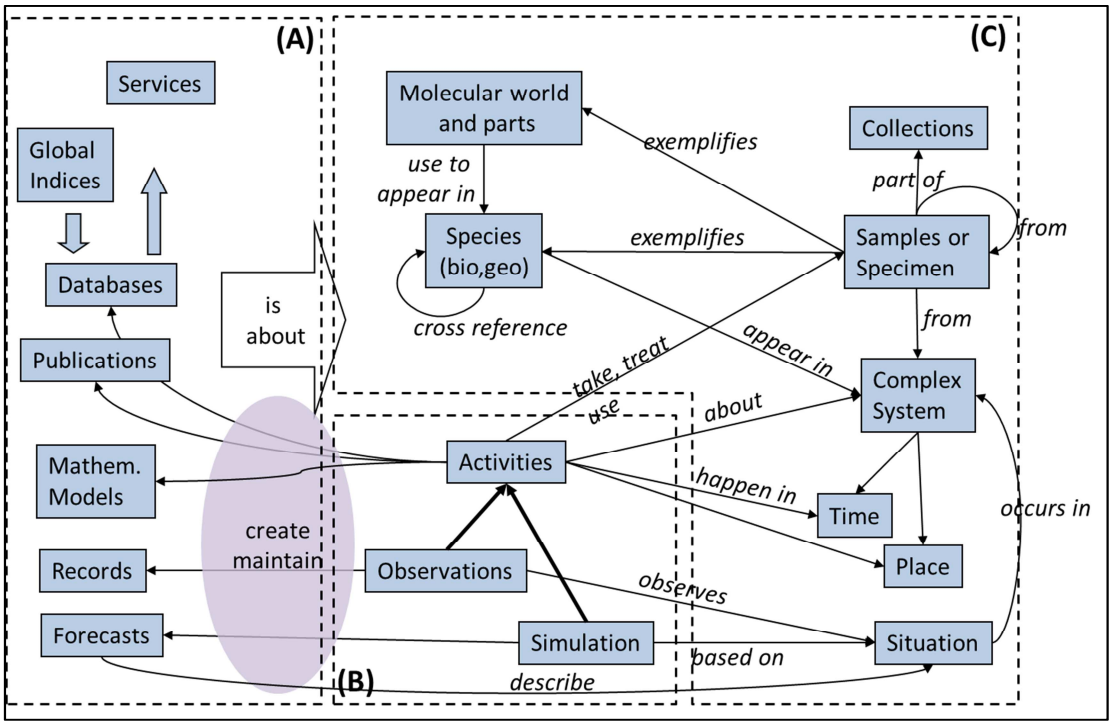

Figure 2 The "Big Picture" conceptually: (A) products of human activities, (B) human activities and (C) core conceptualization of Earth Sciences

The next step is to use this core conceptual model as a guide for identifying what the aforementioned semantic models cover. This provides a top-down visualization of the current state of art of the semantic models that are currently in use in the Earth Science domain which is useful for interoperability purposes. In addition, this enables the identification of overlaps and gaps. From our analysis we have found out that

some of the models described above cover most of it, while others cover only some parts. The following table illustrates the placement of each model in the big picture. The mark symbol ('✓') between a model and a group (e.g. between the ISO-19100 series and group A) denotes that the model is capable for modeling all the aspects from that group, while the star symbol ('★') denotes that only some of the aspects are covered.

	ISO 19100 Series	OGC Standards	CGI Vocabularies	CF Conventiions	GEMET	GeoSciML	OTEG	MOLES	ThIST	VoID	SKOS	Dublin Core	ISO 19127
A	✓	✓		★								★	★
B	✓	✓		✓			✓	✓					✓
C	★	★	★	★	★	★	★	★	★	★			★

Table 3 Coverage of the semantic models with respect to the core conceptual model

STRATEGIES FOR HAVING HARMONIZED METADATA, SEMANTICS AND ONTOLOGIES

According to the Dublin Core Metadata Initiative[14], metadata harmonization is defined as «the ability of two or more systems or components to exchange "combined metadata" conforming to two or more metadata specifications, and to interpret the metadata that has been exchanged in a way that is consistent with the intentions of the creators of the metadata» and as "recipe for harmonization" it recommends adopting a core model with support for machine processable semantics and construct mappings of others standards to the core model.

In general we can identify two main strategies: (a) compliance with standards and (b) adoption of integration/harmonization approaches. Regarding the former, there is adequate standardization (mandated) in the geographical domain. Two major standardization organisations ISO and OGC cooperate together to develop interoperability standards that are increasingly becoming adopted. Below we discuss the second strategy, i.e. approaches for integration/harmonization.

Integration Approaches

The availability of mappings between various models (metadata schemas/ontologies) is crucial for interoperability. They allow building tools and systems for exchanging and integrating information. Specifically they can be exploited for implementing a materialized integration (warehouse), or a virtual integration (mediator) approach.

Materialized Integration relies on a central repository where all data are to be stored, called "Data Warehouse". Mappings are used to extract information from data sources, to transform it to the target model and then to store it at the central repository. It is good practice not to modify extracted information after its transformation except for the use of common identifiers. Rather, any need for updating individual information is covered by requesting source providers to make updated sources available.

There are some important issues that should be taken into account for designing and maintaining a data warehouse. Firstly (designing phase) the information from each source that is going to be used should be

[14] http://dublincore.org/

selected. Specific views over the sources should be chosen in order to be materialized and the global schema employed by the warehouse should be constructed. Next (maintenance phase) issues should be tackled concerning the warehouse initial population by the source data and the update of the data when sources are refreshed. Finally, there are some query processing, storage and indexing issues that should be taken into consideration. The great advantage of materialized integration is the great flexibility in transformation logic, decoupling of the release management of the integrated resource from the management cycles of the sources and the decoupling of access load from the source servers. An example of the process of constructing a warehouse for the biodiversity domain is described in [4].

On the other hand, Virtual Integration approaches do not rely on a central repository but leave the data in the original sources. Mappings are exploited to enable query translation from one model to another. Then data from disparate sources are combined and returned to the user. The mediator (a.k.a. integrator) performs the following actions: first it receives a query formulated in terms of the unified model/schema and decomposes these queries into sub-queries. These queries are addressed to specific data sources. This decomposition is based on the mappings generated between the unified model and the source models, which play an important role in the sub-queries' execution plan optimization. Finally, the sub-queries are sent to the wrappers of the individual sources, which transform them into queries over the sources. The results of these sub-queries are sent back to the mediator. At this point the answers are merged into one and returned to the user. Besides the possibility of asking queries, the mediator has no control over the individual sources. The great advantage (but in some cases disadvantage) of virtual integration is the real-time reflection of source updates in integrated access. The higher complexity of the system and the quality of service demands on the sources is only justified if immediate access to updates is indeed required.

Candidate Core Ontologies and Technologies

Below we describe some possible (current and under development) models that can be used as a basis for achieving harmonization in terms of metadata, semantics and ontologies.

The ISO 19100 series of standards has been the basis for the development of well-known semantic models (including MOLES, GeoSciML and many others). Furthermore ISO is in close cooperation with OGC for the development of emerging standards. The main concern of these standards (ISO 19100 series and OGC Specifications) is to support the interoperability of geo-information systems which are the building blocks of spatial information infrastructures. For achieving interoperability these standards provide a basic conceptualization of: (a) unified data models, (b) well defined interfaces, (c) languages to access and manipulate data based on these interfaces and (d) automatic translation of data and models.

The well-known ISO 19115 standard provides a structure for describing digital geographic metadata. As stated in [5] it is sufficient for capturing enough of the context surrounding the data, however it cannot be used for capturing other important preservation-related metadata as specified in the OAIS reference model [6], such as Representation Information, or Preservation Description Information. In [5] a preservation profile of ISO 19115 has been developed, based on the metadata requirements specified in the OAIS reference model. The main purpose of this profile is to enable capturing preservation-related information about geospatial dataset, while persisting the ability to capture contextual metadata of these datasets using the core ISO 19115 model.

While ISO 19115 is the standard metadata scheme for geographic datasets, its structure is not the most appropriate when dealing with Earth Observation (EO) products: the required ISO19115 metadata elements like title, abstract, contact point, etc. make the description of EO products less efficient, while important attributes for the description of various EO products are missing. To address this problem, a profile of ISO 19156 called the Earth Observation Metadata profile of Observations and Measurements (OGC 10-157) has been developed in the context of the HMA[15] projects. Similarly to ISO 19115 it provides some elements for capturing preservation related information.

[15] http://earth.esa.int/hma/

Alternatively the CIDOC CRM (ISO 21127) [3] provides the concepts to model the complexity of cultural heritage and archaeological data in their semantic context, but also scientific observation, measurement and processing activities. Recent research however has revealed that CIDOC CRM and OGC standards are opaque with respect to the exact relationship of the real extent of the matter of features to given geometries. As a consequence, both standards fail at scale-independent integration of geometric data with data of other semantics. For instance, the given geometry of a detail of a borehole for water sampling may not overlap with the given geometry of the borehole as a whole. To solve this problem, a novel "articulation" of the CIDOC CRM with OGC standards has been produced; a common ontological refinement of both standards capable to explain the missing details as specializations of both models simultaneously. The resulted model, which currently is called CRMgeo [9], is a spatiotemporal extension to the CIDOC CRM expressed in RDF. CRMgeo integrates the CRM with the OGC standard of GeoSPARQL [7] and thus provides the concepts to represent current standardised geo-information in an ontological network with its data, metadata and provenance data. The benefits stemming from such a model is that one could use it for building repositories that host both spatial data and other semantic resources. Moreover, it could be exploited as the schema for information integration, either in a materialized approach, or a virtual approach. The basic service over an integrated repository is the search/query service. In [14] an extension of RDF and SPARQL, namely stRDF, which is capable for representing geospatial data that change over time and a database system which allows querying such data, has been developed.

At this point we should also stress that RDF/S is currently the lingua franca for metadata and semantic descriptions. The adoption of this representation framework has various benefits (exploitation of the tools that have been developed, Linked Data initiative, etc.). According to the W3C definition, the Semantic Web is a collaborative effort, which "provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries." The vision is based on the idea of extending the principles of the Web from documents to data by defining and describing the relations among data on the Web. These meaningful relationships can be established between any named resources in the Web documents by enabling an automatic integration of data.

The Linked Data approach lies at the core of the Semantic Web stack by adopting a small selection of semantic web technologies. This idea is summarized by a set of best practices, introduced by Tim Berners-Lee in [8] for publishing and interlinking structured data on the Web: (1) use URI[16]s as names of things, (2) use HTTP URIs so that people can look up those names, (3) when someone looks up a URI, provide useful information, using the standards (i.e. RDF, SPARQL, etc.) and (4) include links to other URIs, so that they can discover more things. There are already billions of RDF triples which have been published according to those principles. According to many, Linked Data seems to offer a valid approach to address many of the interoperability issues faced in digital preservation providing solutions for data search, virtualization and integration.

To encourage users adopt the Linked Data principles W3C announced the 5-Star-Linked-Data[17] initiative as a best practice for publishing Linked Data. It is a rating system, which allows users ensuring that their data are open-accessible and interlinked. The lowest rating (1 star) is given for datasets which are available on the web (in any format), but with an open license (e.g. a pdf document showing temperature measurements). 2 stars are given for datasets which are accessible on the web in a structured machine-readable format, so anyone can re-use them (e.g. an XLS file with temperature measurements). 3 stars are given to datasets which do not depend on proprietary software (e.g. use CSV instead of XLS). 4 stars are given to datasets which are in the web in the sense that they contain URIs (an RDFa file, so that other resources can point to them). The highest rating (5 stars) is given to data which are linked with other data to provide a description of the context (an RDFa with properties linking to other Linked Data).

[16] Uniform Resource Identifier
[17] http://5stardata.info/

CONCLUSION

In this paper we summarized the results of an analysis of the Earth Science needs, procedures and gaps in terms of the current data preservation and data access policies. The analysis revealed that there is an increasing awareness of the importance of digital preservation in the Earth Science organizations even if there is still often a lack of strategic planning for digital preservation activities. Preservation of scientific research results is in the public interest. Earth Scientists are highly receptive to the adoption of preservation tools. The bridge between user needs and technology must therefore be maintained and reinforced. Moreover it is vital to maintain and provide access to discovery information for data. Custodians of the digital objects have a responsibility to store the data in suitable conditions, ensuring the viability of the original deposited information and migrating the data to new formats when necessary. Additionally several organizations lack a formal data preservation policy even if their definition is an on-going process in nearly all scientific organizations with a different level of maturity among the different stakeholders depending on current funding levels. Finally organizations who do not claim to have a formal data preservation policy nevertheless are aware of the need for preservation and have objectives to this end. Typical data preservation objectives include: (a) the ability to share and exchange data in the long-term, (b) the ability to share and exchange metadata and data products and (c) the facility to minimize the impact of software hardware upgrades on the operation of the archive and on accessibility/usability of the data it contains.

Furthermore in this paper we summarized the results of an analysis of the semantics, metadata and ontologies that are currently in use by earth scientists, in terms of their usability, adoption, configurability, and extensibility across different Earth Science domains. A semantic backbone was described comprising fundamental categories and relationships, providing a kind of high level picture of the relationships between human activities and their targets, as they appear on the highest level of description of all scientific products and services. Moreover, the placement of the various semantic models in that semantic backbone was described. Then we proposed strategies to have harmonized metadata, semantics and ontologies able to satisfy the user needs coping with the different Earth Science domain approaches. In general, one approach is to comply with standards and we have seen that there is adequate standardization (mandated) in the geographical domain: ISO 19156 (for data) and ISO 19115 (for metadata). The second strategy is to adopt integration/harmonization approaches which rely on mappings. Moreover we stressed the potential of the Semantic Web and Linked Data trend. According to many, Linked Data seems to offer a valid approach to address many of the interoperability issues faced in digital preservation providing solutions for data search, virtualization and integration. In the domain of cultural heritage, for example, Semantic Web technologies have been proposed as a promising approach for making multi-format, multi-topical, multi-lingual, multi-cultural and multi-target content mutually interoperable (at syntactic but especially at semantic and organizational levels) so that it can be searched, linked and published in a harmonized way across the boundaries of the datasets and data silos [13].

REFERENCES

[1]- SCIDIP-ES. Science Data Infrastructure for Preservation – Earth Science. Deliverable on WP33, D33.1- Earth Science common preservation policies and data access policies analysis, July 2013.

[2]- SCIDIP-ES. Science Data Infrastructure for Preservation – Earth Science. Deliverable on WP33, D33.2-Earth Science metadata, semantics and ontologies harmonization report, June 2013.

[3]- Partick Le Boeuf, Martin Doerr, Christian Emil Ore and Stephen Stead. “Definition of the CIDOC Conceptual Reference Model”, November 2012. http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.pdf

[4]- Yannis Tzitzikas, Carlo Alloca, Chryssoula Bekiari, Yannis Marketakis, Pavlos Fafalios, Martin Doerr, Nikos Minadakis, Theodore Patkos and Leonardo Candela, “Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology”, 7th Metadata and Semantics Research Conference, MTSR 2013, Thessaloniki, Greece, November 2013.

[5]- Arif Shaon and Andrew Woolf. “Long-term Preservation for Spatial Data Infrastructures: a Metadata Framework and Geo-portal Implementation”. D-Lib Magazine, vol. 17. October 2011.

[6]- International Organization for Standardization. OAIS: Open Archival Information System – Reference Model Ref. No ISO 14721:2003.

[7]- Open Geospatial Consortium. “OGC GeoSPARQL- A geographic query language for RDF data”. 2012. http://www.opengeospatial.org/standards/geosparql

[8]- Linked Data: Design Issues. Tim Berners-Lee. W3C, Last Update: 2009-06-18, http://www.w3c.org/DesignIssues/LinkedData.html

[9]- Martin Doerr and Gerald Hiebel. CRMgeo: Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement. Technical Report – ICS-FORTH/TR-435, http://www.ics.forth.gr/tech-reports/2013/2013.TR435_CRMgeo_CIDOC_CRM_GeoSPARQL.pdf

[10]- Spiros Ventouras, Bryan Lawrence, Andrew Woolf and Simon Cox. “MOLES Information Model”. In EGU General Assembly Conferene Abstracts, vol. 12, p. 5080, 2012.

[11]- Alistair Miles and Sean Bechhofer. “SKOS Simple Knowledge Organization System reference”. Technical report, W3C, 2009.

[12]- Dublin Core Metadata Initiative and others. “Dublin Core metadata element set, version 1.1”, 2008.

[13]- APARSEN. Alliance for Permanent Access to the Records of Science Network. Deliverable on WP25, D25.1-Interoperability Objectives and Approaches (February, 2013).

[14]- Manolis Koubarakis, Charalambos Kontoes, Stefan Manegold, Mihai Datcu, Ugi Di Giammatteo, Eva Klien. TELEIOS: A Database-Powered Virtual Earth Observatory. In proccedings of the 38th International Conference on Very Large Data Bases, Istanbul, Turkey, August 2012.