SCIENCE DATA INFRASTRUCTURE FOR PRESERVATION - EARTH SCIENCE

Mirko Albani[1] Fulvio Marelli[1,] David Giaretta[2] Arif Shaon[3,]

[1]European Space Agency ESA-ESRIN, Italy
[2]Alliance for Permanent Access, The Netherlands
[3]Science and Technology Facility Council, United Kingdom

KEY WORDS: Long term data preservation, Earth Science, Preservation Policies and techniques, Earth Observation

ABSTRACT:

The proper preservation of both current and historical scientific data will underpin a multitude of ecological, economic and political decisions in the future of our society. The SCIDIP-ES project addresses the long-term persistent storage, access and management needs of scientific data by providing preservation infrastructure services. Taking exemplars from the Earth Science domain we highlight the key preservation challenges and barriers to be overcome by the SCIDIP-ES infrastructure. SCIDIP-ES augments existing science data e-infrastructures by adding specific services and toolkits, which implement core preservation concepts, thus guaranteeing the long-term access to data assets across and beyond their designated communities.

1. INTRODUCTION

1.1 The Challenge

Climate change, environmental degradation and ecological sustainability are amongst the most vital aspects that need to be understood and managed today and in future. Understanding these challenges involves the complex analysis of environmental information, such as Earth Science data to inform government policy and practical implementation in areas (e.g. climate change, water management, health and agriculture) that underpin the stability of existing socio-economic and political systems. Thus there is a need to preserve a flood of Earth Science (ES) data and, more importantly, the associated knowledge to ensure its meaningful long term exploitation. Moreover, certain environmental analysis, like those supporting the long-term climate change variables measurement, requires historical data records to be periodically reprocessed to conform to the latest revisions of scientific understanding and modelling techniques. This in turn requires access to and understanding of the original processing, including scientific papers, algorithm documentation, processing sources code, calibration tables, databases and ancillary datasets.

To maximise the value of ES data, its usage should not be limited to the domain scientists who originally produced it. ES data as a "research asset" should be made available to all experts of the scientific community both now and in the future. The ability to re-purpose existing ES data could cross-fertilise research in other scientific domains. For example, if epidemiologists can correctly interpret environmental data encoded in an unfamiliar format, the additional knowledge may assist them with understanding patterns of disease transmission.

Unfortunately getting access to all the necessary data and metadata is a serious problem; often the data are not available, accessible or simply cannot be used since relevant information explaining how to do so or the necessary tools, algorithms, or other pieces of the puzzle are missing. Moreover the ES data owners are dealing with the preservation and access of their own data and this is often carried out on a case by case basis without established cross-domain approaches, procedures and tools.

The SCIence Data Infrastructure for Preservation – Earth Science (SCIDIP-ES) project [1] is developing services and toolkits which can help any organisation but the prime focus in this project is to show their use in ES organisations working with non-ES organisations concerned with data preservation to confirm the wide effectiveness in helping to improve, and reduce the cost of, the way in which they preserve their ES data holdings. In parallel, the project is will produce harmonized models for Earth Science data preservation policies, technologies, semantics and ontologies. This is carried out in tied coordination with the work already undertaken by the Long Term Data Preservation Working Group, which has developed Guidelines for Earth Observation data preservation. The goal is to harmonize and extend the model to the Earth Science wider sector.

2. BARRIERS AND CHALLENGES OF EARTH SCIENCE DATA PRESERVATION TITLE

Here, we discuss some of the key challenges of preserving ES data considered by the SCIDIP-ES project. We have identified these challenges based on the results of a series of surveys conducted by SCIDIP-ES on various aspects of preserving ES data, as well as related external materials, such as the PARSE.Insight case studies [2] on the preservation of Earth Observation (EO) data. Notably, some of the issues outlined here are also relevant beyond the ES and EO domains to the wider data preservation problem.

2.1 Ensuring Intelligibility an (Re-) Usability of Data

A frequently repeated mantra for digital preservation activities is "emulate or migrate", which is also pertinent to the ES data. However, while these activities may be sufficient for rendered objects, such as documents or images, they are not enough for other types of digital objects. In addition, there is a need to capture Representation Information (RepInfo) - a notion defined by the widely adopted ISO standard [3] Open Archival Information Systems (OAIS) Reference Model [4] to represent the information needed to access, understand, render and (re)use digital objects. The key aspects of RepInfo needed to ensure continued intelligibility and usability of data include Semantic

Representation Information (i.e. intended meaning and surrounding context of data) and the identification of a Designated Community (consumer of the data).

2.2 Designing Cost Effective Preservation

Long-term preservation archives and repositories must plan responses to changes and risks of changes in an appropriate and cost-effective way. As discussed above there are many different types of preservation action/strategy which are equally valid and need to be considered when a preservation solution is formulated for a data collection. Archives need to be aware of, characterise and describe the main types of preservation action available to an archivist. They also need to appreciate the effect each type of action has upon a network of RepInfo, the risks, available modes of stabilisation as well as cost and benefits. Hence there is a need for tools to help to evaluate and balance costs and risks in a network of RepInfo In addition, they need to consider how more than one type of strategy can be employed as alternates in order to create the optimal balance of risk and usability of a preservation solution.

2.3 Reacting to changes in preservation requirements

As mentioned above, long-term data archives need to be able handle changes in preservation requirements by re-strategizing when needed. It is well understood that hardware and software become unavailable but also the semantics of specific terminology change and the knowledge base of the Designated Community, as chosen by a repository, changes. All these changes must be countered if we are to preserve our digitally encoded information. Yet how can any single repository know of these changes? Significant effort (e.g. the preservation watch service of the SCAPE project [5]) is being put into technology watches for document and image format changes. It is more difficult for a repository to watch for all possible changes, such as in terminological changes across a multitude of scientific disciplines, and to understand the ramifications of such changes. From this perspective, there is a need for services to spread the knowledge about such changes, or the risk of such changes, and the implications of such changes.

2.4 Maintaining Authenticity

In general, any process and transformation could have side effects on digital data and corrupt its usability and integrity of the information being preserved. Therefore, authenticity requires more than just digital digests (e.g. checksum) – because these cannot by themselves guarantee that the data has not been altered, by accident or on purpose, by those in charge of the data and digests. Moreover the data may have been transformed from one form to another over time for a variety of reasons – the bit sequences and therefore the digests will change. More generally authenticity is not a yes/no issue – such as "does the digest match or not" – but rather a degree of authenticity judged on the basis of technical and non-technical evidence.

2.5 Supporting Practical Business Models for Data Preservation

Preservation of data requires resources and long term commitments; an important aspect is therefore the need for business models in order to build business cases for well identified "research assets" which can justify their continued funding. At the same time the costs of preservation must also be reduced by avoiding unnecessary duplication of effort and wasting of resources, including energy. For instance, it may be financially more viable to turn an existing storage system into a preservation archive by integrating preservation services and tools into the existing system than to create a separate preservation archive. However, no organization can guarantee its ability to fund this storage and those responsible for the data will change over time. Long-term sustainability requires more than good intentions. It requires funding, and the recognition that the costs must be shared wherever possible. It also requires one to be realistic and recognize that no one repository can guarantee its existence forever; one must be prepared to hand over the digital holdings in a chain of preservation which is only as strong as its weakest link – and the hand-over from one link to the next must be easy and flawless. This hand-over is not just transfer of the bits but also the information which is normally held tacitly in the head of the data manager or embedded in the host data management system. We envisage that suitable and efficient services and tools can help prepare repositories for the hand-over process and moreover share the results and experience with the wider preservation community.

3. THE SCIDIP-ES PROJECT

The SCIDIP-ES consortium puts together a group of partners, which covers from two different perspectives the theme of digital data preservation.
On one side is constituted by earth science data creators, curators and providers. It is constituted by three main European Space Agencies – such as ESA, DLR and CNES – plus data curators and providers belonging to a wider Earth Science community, including STFC, NERC, INGV and ISPRA.
The consortium also includes partners coming from a consolidated path of digital preservation research projects: starting from the Alliance for Permanent Access, it includes technical, commercial and academic partners involved in the last decade on digital preservation projects such as CASPAR [6], Parse.INSIGHT and SCHAMAAN, etc. These include industrial partners – ACS, Engineering, ICT, GIM, CapGemini – and partners belonging to the academic world: JUB, UTV, Forth, FTK.

The project's aims

- Upgrade CASPAR prototype components into scalable, robust e-infrastructure components to support digital preservation of all types of digital objects.
- Harmonize policies, ontologies and semantics for data preservation and future use.
- Set-up a European framework for the long term preservation of Earth Science data

SCIDIP-ES Services and Toolkits

Preservation requires, besides keeping bits, ensuring the information encoded in a digital object continues to be usable, and there is evidence that the digital object is what it is claimed to be. The SCIDIP-ES services and toolkits help this to be done. To ensure these services have a user base after the project we must ensure that the services are tuned to Earth Science repositories' and users' existing systems, showing that at least some consortium data – new as well as old – is usable where it is unfamiliar. The services must be shown to be usable by and customisable for other communities and must be implemented in a way, which allows them to be supported, by the end of the project. All of the tools and services must be designed to be

customisable so that they can fit into existing (and we hope near future) systems and applications. The "core" of each of the services, which can be customised for a variety of domains and systems, must be easily maintainable and supportable after the end of the project. The toolkits will be run on various peoples' desktops whereas the services themselves could be run by a single organisation, shared by everyone in that organisation; alternatively they could be run by a variety of organisations, sharing the services between each other or even with outside users.

Harmonization of Metadata, Semantics and Ontologies: The SCIDIP-ES project, after performing a survey on the current metadata, semantics and ontologies available for Earth Science data and on the current related initiatives, will define and validate an appropriate strategy to have harmonized metadata, semantics and ontologies able to satisfy user needs coping with the different Earth Science domains approaches. The strategy consists for example in the definition of a common ontology targeting at covering all, starting from a subset, the possible Earth Science applications domains and data categories or, more likely, at demonstrating the viability of a "semantic mediated access across domains" approach able to make the different available ontologies communicate between each others. For what concerns the metadata harmonization, we will analyse and extend the HMA approach and results to other data categories exploiting the experience of the consortium members. We will moreover harmonise the information models used for earth observation data with the ones used for insitu, airborne, balloons, etc. This activity shall address the harmonisation of the data in point via the analysis of recommended standards and best practices in the field and so propose an efficient costeffective methodology for applying such harmonisation. In particular an harmonized information model for all kinds of raster data occurring in the Earth Sciences will be developed. Examples include 1-D in situ sensor data, 2-D EO imagery, 3-D image time series (x/y/t) and exploration data (x/y/z), and 4D climate and ocean data (x/y/z/t). Based on a common raster query language such data can be integrated seamlessly across all Earth Science domains, enabling for unified cross-domain access (e.g., integrating climate data with GIS data).

ES Data Preservation Policies:

After performing a survey on the current preservation policies and guidelines available for Earth Science data, we will define, starting from the outcomes of the survey, common data preservation policies applicable to all Earth Science data categories in order to pursue harmonization of the preservation approach of the different data producers and providers to the maximum extent within and among the different data categories with the goal also to minimize costs and maximizing interoperability and synergies. The common policies will also contain the definition, to the best today understanding, of the knowledge associated to each data category to be preserved in the different data domains to satisfy today and future user needs. The definition and application of these policies will help to create a collaborative framework among Core Earth Science data user communities (e.g., land, ocean, atmosphere..) and data owners in Europe. The harmonization of rights and Intellectual property frameworks for the access to Earth Science data and associated knowledge will also be analysed and addressed in line with EC directives and International agreements such as INSPIRE and the "GEOGEOSS Data Sharing Principles" with the goal to pursue harmonization and simplification of access for users. The possibility to define and propose new data access policies for example for some subsets of Earth Science data (e.g. Earth Observation Historical data) will also be considered.

Earth Science LTDP Framework governance model and architecture

Impact analysis on the current infrastructure of the different initiative participants in the different data domains will be performed in light of the Earth Science Infrastructure principles. The architecture of a European Infrastructure, based on the upgrade and federation of existing components and on the integration of the generic services developed in the project will be defined. In addition to technical infrastructure and capabilities, the long-term management of Earth Science data requires organizational sustainability to provide continuing stewardship to address the risks to scientific data and support their use by future communities. Providing sustainable infrastructure for the preservation of scientific data requires organizational commitments, capacity, structures and plans for data stewardship that are consistent with the missions of the organizations that accept the responsibility to serve in data stewardship roles. Alternative approaches to attaining organizational sustainability for interdisciplinary human dimensions and polar data are discussed in terms of recent recommendations for organizational sustainability to foster digital preservation. To this end SCIDIPES will also define the governance and organization model of the ES infrastructure with the goal to achieve sustainability in the long term, according to the sustainability models adopted for example in the ESFRI projects, and to pursue a maximisation of the open access to data for users respecting individual provider's data policies where necessary.

General Approach

The project approach is informed by the recently published HLEG report, which calls for an international framework for a Collaborative Data Infrastructure. One aspect of their vision was that "Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data and they can evaluate the degree to which the data can be trusted". The SCIDIP-ES team will take address of these in the following ways:

- By working closely with real users, in particular but not limited to the Earth Science domain, and building what they require, thus ensuring their adoption of the infrastructure services.
- By ensuring there is an effective governance and maintenance of the services from the start, and by not trying to impose a top-down system, the consortium will help to ensure that there is an infrastructure which is not too complex to work.
- By addressing disciplinary and cross-disciplinary strategies for metadata definition we will ensure that data can be re-used.
- By applying the subsidiarity principle – so we do not to appear to tread on researchers' toes – and taking advantage of the growing need for researchers to use data from outside their own discipline, we will overcome lack of willingness of projects/funders/nations to take part and use the infrastructure services.

4. CONCLUSIONS

The proven generic services developed in SCIDIP-ES will be tailored to the Earth Science domain specific needs. Harmonization of rights and Intellectual property frameworks for the access to Earth Science data and associated knowledge will also be analysed and addressed in line with EC directives such as INSPIRE and the "GEO-GEOSS Data Sharing Principles". The goal is to achieve sustainability in the long term, according to the sustainability models adopted for example in the ESFRI projects to facilitate access to data for users, while respecting data providers' policies where necessary. As such the Initiative will pave the way for the establishment of the core of a persistent and robust Earth Science infrastructure in Europe, starting from the infrastructure of the partners involved in the SCIDIP-ES consortium, able to respond to the needs of data-intensive science applications addressing for example environmental, climate change (for very long term data analysis integrating historical data taken by historical / scattered instrumentations with recent, more sophisticated, sensors) and disaster monitoring (immediate response to unknown situations for generating specialised operation information). ESA experience in the set up of the GMES Space Component and Coordinated Data Access System (GSCDA) will be a fundamental and unique skill able to guarantee the success of the SCIDIP-ES initiative.

5. REFERENCES

[1] www.scidip-es.eu

[2] www.parse-insight.eu

[3] ISO 14721:2003 - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

[4] http://www.mendeley.com/research/reference-model-for-service-oriented-architectures/

[5] The SCalable Preservation Environment (SCAPE) project - http://www.scape-project.eu/

[6] www.casparpreserves.eu