Towards a Long-term Preservation Infrastructure for Earth Science Data

Arif Shaon, David Giaretta[1], Esther Conway, Brian Matthews, Shirley Crompton
Science and Technology Facilities Council
Rutherford Appleton Laboratory
Didcot, UK
{arif.shaon, david.giaretta, esther.conway, brian.matthews, shirley.crompton}@stfc.ac.uk

[1]also Alliance for Permanent Access,
director@alliancepermanentaccess.org

Jinsongdi Yu,
Jacobs University
Campus Ring 1, 28759 Bremen
Germany
j.yu@jacobs-university.de

Fulvio Marelli
European Space Research Institute/European Space Agency
Via Galileo Galilei, Casella Postale 64
Frascati (Rome), Italy
fulvio.marelli@esa.int

Ugo Di Giammatteo
Advanced Computer Systems
Via della Bufalotta 378,
Rome, Italy
udig@acsys.it

Yannis Marketakis, Yannis Tzitzikas
Foundation for Research and Technology - Hellas (FORTH)
Institute of Computer Science
N. Plastira 100
Heraklion, Crete, Greece
{marketak, tzitzik}@ics.forth.gr

Raffaele Guarino
Capgemini
Rome, Italy
raffaele.guarino@capgemini.com

Holger Brocks
InConTec GmbH
Kirschenallee 7 96152 Burghaslach, Germany
holger.brocks@incontec.de

Felix Engel
FTK Association Research Institute for Telecommunications and cooperation
Martin-Schmeißer-Weg 4,
Dortmund, Germany
fengel@ftk.de

ABSTRACT

The effective preservation of both current and historical scientific data will underpin a multitude of ecological, economic and political decisions that shape the future of our society. The SCIDIP-ES project addresses the long-term preservation of the knowledge encoded in scientific data by providing preservation e-infrastructure services which support the persistent storage, access and management needs. Using 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 and exploitation of data assets across and beyond their designated communities.

Keywords

digital preservation, e-infrastructure, earth science, services.

1. INTRODUCTION

Climate change, environmental degradation and ecological sustainability are amongst the most vital issues that need to be understood and managed today and in future. Understanding these challenges involves the complex analysis of environmental information, including 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 [9]. 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 analyses, such as 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 of the 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 the following we describe how these services and tools are used to help to overcome some of the aforementioned problems faced by both the curators and the users of ES data, but it should be remembered that they are designed for much wider applicability.

In this paper, we discuss the key technical challenges and barriers of long-term ES data preservation that the SCIDIP-ES project is aiming to address. In addition, we highlight some examples

[1] The SCIDIP-ES project - http://www.scidip-es.eu/

gathered from the ES community during the first year of the project and present the SCIDIP–ES services and toolkits as solution to these community generated requirements.

2. BARRIERS AND CHALLENGES OF ES DATA PRESERVATION

The SCIDIP-ES project identified the following challenges based on the results of a series of surveys on various aspects of preserving ES data within the project, as well as related external materials, such as the PARSE.Insight case studies on the preservation of Earth Observation (EO) data [10]. 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 and (Re-) Usability of Data

A frequently repeated mantra for digital preservation activities is "emulate or migrate". 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[2] Open Archival Information Systems (OAIS) Reference Model [1] 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).

Take for example some fairly simple tabular scientific data in an Excel spreadsheet. This can be easily migrated (or more accurately "transformed" in OAIS terms) to a comma-separated values (CSV) file. However if the semantics, such as the meaning of the columns and the units of the measurements is not recognised as important and preserved then the data will become meaningless and scientifically unusable. The problem is even more important for complex scientific data. Emulation to enable the continued use of the software used to handle the digital objects may be adequate for rendering these objects or re-performing previous operations. However, to combine the preserved data with newer scientific data will, in general, not be possible. For example, one may use an emulator to continue using the Excel software which has the semantics of what the columns mean encoded in its formulae, but one will not be able to combine this data with newer data, for example in NetCDF format[3] which is a commonly used ES data format. Since emulators are a type of RepInfo, one can re-state the mantra as "collect RepInfo or Transform".

This means that a key problem we need to address is – *how does a repository create or collect enough RepInfo?* It is difficult enough to deal with the complex dependencies of an ES data format like NetCDF; when one then looks at the multitude of ES and other scientific formats, each of which may have a plethora of associated semantic RepInfo (thus forming a tree or network of RepInfo dependencies), the problem explodes! In general, an archive may, depending on its data holdings, need various such networks - both individual and related. Hence, there is a need for a service and tools to help spread the load in creating and managing RepInfo networks in a preservation archive or repository.

2.2 Designing a cost effective preservation solution

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 RepInfo Network, the risks, available modes of stabilisation as well as cost and benefits. Hence, there is a need for tools to help evaluate and balance costs and risks in a RepInfo network. 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-strategising 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[4]) is being put into technology watches for document and image format changes. It is more difficult for a single repository to monitor 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, risk and implications of such changes.

2.4 Maintaining Authenticity

It is important to guarantee within an archive that digital data is managed and maintained through proper tools by applying suitable plans in order to ensure the "authenticity" of the data. In the OAIS model, authenticity of digital object is defined as *"the degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence."* [1]

In general, any process and transformation could have side effects on digital data and corrupt the 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

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

[3] NetCDF - http://www.unidata.ucar.edu/software/netcdf/docs/

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

technical and non-technical evidence. In effect, this involves capturing and evaluating that evidence as it is generated in many different ways over an extended time. Performing these tasks manually is likely to be laborious and even erroneous. This underlines the need for suitable tooling to facilitate capturing and evaluating the evidence needed to guarantee authenticity of data in a digital preservation archive.

2.5 Supporting Practical Business Models for Data Preservation

Preservation of data requires resources and long term commitments; therefore we need practical business models in order to build business cases for well identified "research assets" to 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 organisation 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 recognise that no one repository can guarantee its existence forever; one must be prepared to hand over the digital holdings in a chain of preservation that 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. KEY USE CASES CONSIDERED IN SCIDIP-ES

The SCIDIP-ES project has defined the following three high level use cases to represent the main challenges of long-term preservation of ES data discussed above.

- **Preservation Archive Creation:** identifying what kind of information should be properly preserved for future use, by an identified Designated Community (DC) and the correct procedures needed to implement it. For existing archival systems, this would also need to address the efficient integration of preservation processes within the underlying system architecture.
- **Archived Data Access**: to add value to the preserved data, what kind of enhanced information could be provided to current and future consumers? In particular how can the repository enable a broader set of users to understand and use its data, e.g. to build a broader ES community, beyond the initial DC.
- **Archive Change/Evolution**: how to preserve data against changes in related technology (e.g. hardware, software) and in the designated community (data producer, data preserver, data consumer, the communities and organization involved in the information's creation and initial use).

In this section, we describe the first high level use case – Preservation Archive Creation– with a specific focus on the ESA ENVISAT MERIS dataset[5]. This is because the Archive Creation/Enhancement use case is the milestone on which the following use cases are built.

3.1 Preservation Archive Creation

We have defined the following logical model as a guideline to structure the archive definition phase workflow (Figure 1).

Figure 1. Phases of Preservation Archive Creation

3.1.1 Define the Preservation Objective

A preservation objective defines the minimum level and type of reuse which an archive wishes to maintain for its user community. Typical this would cover areas such data processing, visualization, analysis and interpretation of data. For example, MERIS has provided ten years of detailed observations of land, atmosphere and oceans. The Objective is to preserve ESA's MERIS data package to maintain its time series, accessible and usable by different scientific user communities for 50 years. The minimum guaranteed level of preservation is the storage/archiving of the **ESA MERIS N1 File Level 0** (L0) and **Level 1 (L1).** The L0 data is the lowest level product and derived from MERIS. It is the satellite raw data which has been simply reformatted and time ordered in a computer readable format. L1 is derived from L0 data and both use the N1[6] file format. L1 data, among other processes, is geo-located, calibrated and separated from auxiliary data. We focus in this section on preservation of L0 and will discuss the preservation of L1 data in subsequent sections.

3.1.2 Definition of the Designated Communities

The definition of the DC should specify the skills, resources and knowledge base a community has access to. DC description must have sufficient detail to permit meaningful decisions to be made regarding information requirements for effective re-use of the data. In the MERIS case, the DCs (both archive and user community) include:

- ESA staff – with full specific knowledge of ENVISAT datasets management.
- Principal Investigator (PI) - working on Earth topics such as Agriculture, Atmosphere, land, Natural disaster, Ocean, etc. They know the ENVISAT data scientific value but don't have the skills to manage it.
- University Students - they are learning ENVISAT data and need to fully understand and use it.

[5] ENVISAT Meris Instrument description and access to data can be found at https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/meris

[6] The N1 File Structure - http://www.noc.soton.ac.uk/bilko/envisat/tutorial/html/t0110.html#sh2

3.1.3 Preserved Dataset Content Definition

Once the objectives and communities have been identified and described, an archive should be in a position to determine the full set of information required to achieve an objective for this community. To allow processing, visualization, analysis and interpretation of ESA MERIS data and the correct utilization by anyone with basic knowledge of the EO domain, the Archive must contain comprehensive information about:

- Science Data Records: raw data, L0 and L1 data, browse images, ancillary data, auxiliary data , Calibration and Validation data
- Processing software and databases: L0 consolidation software, instrument processing software, quality control software, data visualization tools
- Mission Documentation

3.1.4 Create Inventory

The next stage is to appraise each of the information objects in terms of physical state, location and ownership. The resulting inventory should include details of each of the pieces of Information, its Location, Physical State and associated Intellectual Property Rights (IPR). For example, the MERIS inventory would contain MERIS processing software and databases including:

- L0 consolidation software (mission dependent) described in the mission products description document. This document is available and is the IP of ESA.
- The Basic ENVISAT Toolbox developed to facilitate the utilization, viewing and processing of ENVISAT MERIS data along with the associated GNU public license
- The Java Virtual Machine required to run the ENVISAT Toolbox; Oracle owns several aspects of JAVA related IP.

3.1.5 Perform Risk Assessment

There may be a number of key risks associated with the MERIS data as described in the following categories and examples:

Technical Risk: software for processing the MERIS data (e.g. BEAM software) run with specific libraries (e.g. JVM1.5). Thus, it is also necessary to preserve such information so that the whole chain of soft/hardware dependencies could be evaluated.

Organizational Risks: ESA may decide to store copies of the MERIS data in different geographical locations to safeguard the archive from external hazards like floods and other natural disasters or technological hazards, etc.

IPR related Risks: As a research organization, ESA encourages, protects and licenses innovations or original works resulting from its activities. The MERIS data is protected according to the ESA IPR guidelines[7]. The need is to ensure that IPR or licences related to data, software (e.g. BEAM) and libraries (e.g. Java 1.5) are assessed for potential breaches.

Resourcing Risks: The preservation plan exists on the basis that funding and skills to support the data archive will be available for a defined time period. Should any of this change, the plan will need to be adapted.

3.1.6 Preservation Planning and Risk Monitoring

Preservation planning is the process which designs the long term research asset to be preserved within an Archival Information Package (AIP). AIP conceptually contains all the information required to ensure the long term usability of digitally encoded information. The cost, benefits and risk burden acceptable to an archive will determine the optimal preservation action to adopt. Preservation actions for construction and maintenance of the AIP take one of the following forms: *Risk Acceptance and Monitoring (referencing), Software Preservation or Description and Transformation*.

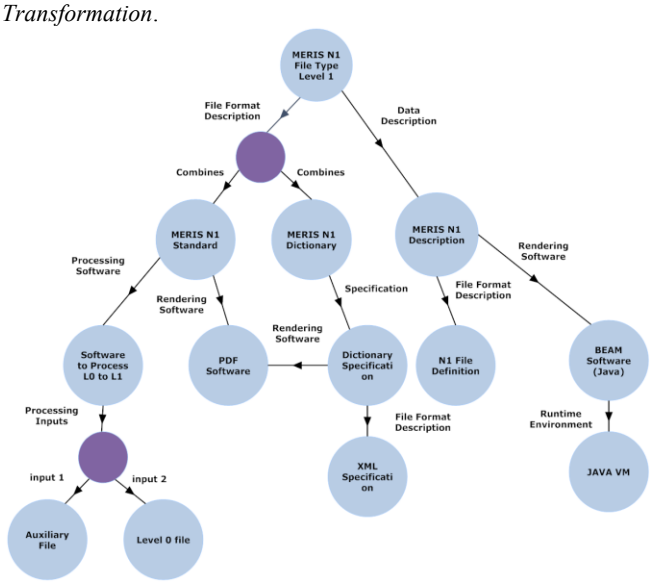

Figure 2. Network of RepInfo for ENVISAT MERIS data

Using the notion of Preservation Network Model described in Section 4.1.4, we designed a network of RepInfo (Figure 2) for the MERIS L1 example. This defines the whole chain of dependencies required to preserve its data and the associated knowledge to interpret it. As no preservation solution is permanent or necessarily stands the test of time, AIPs must be monitored for stability and suitability. To achieve this, the accepted risks/dependencies within the preservation network as well as the preservation objective and DC description must be recorded and monitored.

4. SCIDIP-ES PRESERVATION INFRASTRUCTURE

To address the long-term preservation challenges of the ES data (Section 2), in SCIDIP-ES, we aim to put in place an e-infrastructure consisting of various services and toolkits to facilitate long-term data preservation and usability. In essence, we combine a top-down, data centric view, using a proven design for generic infrastructure services to enable persistent storage, access and management, with a bottom-up, user-centric view, based on requirements from the ES community. The former comes from leading research projects in digital preservation, in particular CASPAR. The latter is from the developing European Framework on Long Term Data Preservation (LDTP, coordinated by ESA) for Earth Observation data.

[7] ESA Intellectual Property Rights - http://www.esa.int/esaMI/Intellectual_Property_Rights/index.html

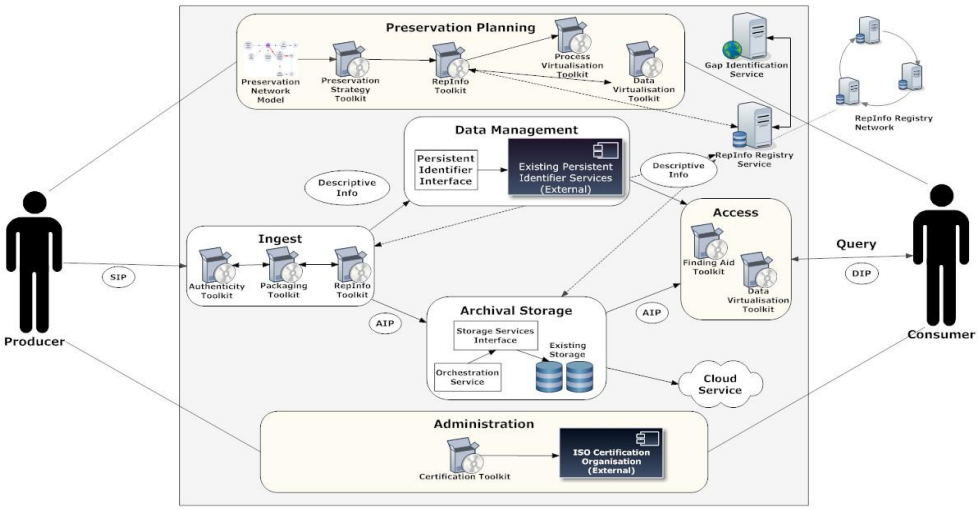

Figure 3. Overview of the services and toolkits within the SCIDIP-ES preservation infrastructure

4.1 SCIDIP-ES Preservation Services and Toolkits

To ensure consistency and interoperability, we use the OAIS Reference Model to underpin the definitions of the services and toolkits of the SCIDIP-ES preservation infrastructure. While the infrastructure is intended to cover the full preservation workflow defined in the OAIS model, the specific areas focused on are those connected with the construction of AIPs. As discussed earlier in the paper (Section 2.5), no organisation can be expected to look after a piece of data forever but rather that it can hand on its holdings to the next in the chain of preservation. Such a process can be hindered by lack of clear understanding of tacit dependencies and knowledge, and insufficient time available during the hand-over to capture these. Creation of an AIP ensures that these are made explicit well before they are needed, and so any future hand-over can be smooth and complete.

The SCIDIP-ES preservation infrastructure consists of the following services and toolkits (as shown in Figure 3) that have been defined to support both the data and user centric views that we have adopted. It enables the ES repositories to effectively address the challenges of preserving ES data mentioned in Section 2. A logical overview of these services and how they support the stages of the OAIS reference model is given in section 3.

4.1.1 RepInfo Registry Service

The RepInfo Registry Service is essentially a web-service based repository that is used to store, query, retrieve and manage the RepInfo needed to enable access, understanding and (re-)use of a digital object over the long-term. The RepInfo provided by the Registry Service can cover the structure of the digital object (format, headers, footers, instrument measures, annotations, fixed parts, variable parts, etc.), the semantics of that digital object (semantics, auxiliary information, usage information), and other information (e.g. rendering information which describes what additional software can be used to display/process/edit the digital object).

The Registry will also enable users to navigate a RepInfo network to explore the knowledge represented (e.g. a satellite image is linked to the image sensor description, which is linked to the satellite mission description, etc.).

4.1.2 The RepInfo Toolkit

The RepInfo toolkit provides a user-friendly GUI to the Registry to enable various components to interact efficiently with it. For example, the ingest and planning of the preservation life cycle in an archive. It also provides users with a set of tools to create the RepInfo required for specific digital objects. Some sub-components of this toolkit are aimed at describing the data in more "virtualised" terms which can help integrate data into other software.

4.1.3 Gap Identification Service

As underlined in the OAIS model, there is a need for services that help archivists in checking whether the archived digital artefacts remain understandable, and to identify hazards and the consequences of probable losses or obsolescence risks. In SCIDIP-ES, we have defined the Gap Identification Service (GIS) to facilitate such assessments of intelligibility of digital objects by identifying "gaps" in the corresponding RepInfo Network in the RepInfo Registry. In essence, this service is inspired by a model that consists of the notions of module, dependency and profile as discussed in [3]. If applied to digital objects, a module can be a software/hardware component or even a part of the knowledge base expressed either formally or informally, explicitly or tacitly, that we want to preserve. The dependency is captured in the logical links in meaning between modules. In addition, a module may require the availability of other modules in order to function, be understood or managed (e.g. a network of RepInfo). A profile is the set of modules that are assumed to be known (available or intelligible) by a user (or community of users), so this is an explicit representation of the concept of Designated Community Knowledge Base (KB). Utilising this model, the GIS is able to check whether a digital object (module) is *intelligible by a community*, and to compute the *intelligibility gap* of a digital object.

In an archive, the GIS can be used in the preservation planning process to evaluate the current knowledge base of the designated community as well as future review(s) of the plans by analysing changes in the related knowledge base. It can also be used for deriving DC-aware AIPs or DIP (dissemination information packages).

4.1.4 Preservation Strategy Toolkit

There are a number of basic strategies for preserving digitally encoded information. Besides describing the data using RepInfo,

one could transform the data into a different format or emulate the essential software to access the preserved information. The Preservation Strategy Toolkit helps repositories decide which technique to use, balancing costs against efficacy for given specific preservation objectives.

The toolkit uses the Preservation Network Model (PNM - see Figure 2 for an example), which was developed within the CASPAR project in order to represent the output of a preservation analysis conducted for a digital object to be preserved in a preservation archive or repository [4]. The preservation analysis of a digital object enables identification and assessment of the risks associated with its dependencies on other entities. The output of this type of analysis underpins the formulation of a suitable preservation strategy to be adopted by an archive; taking into account the preservation aims, related risk tolerance level, preservation policies and other requirements. The PNM can be used to articulate the result of preservation analysis as a network of related objects along with the preservation decisions associated with the relationships between the objects.

In an OAIS-compliant archive, the use of this toolkit would be a part of the **preservation planning** process/stage (see section 3.1.6), where the toolkit would also need to interact with the RepInfo toolkit to query and retrieve existing RepInfo records and/or create new ones as determined by the planning process.

4.1.5 Authenticity Toolkit

The Authenticity Toolkit is used to capture appropriate evidence of the authenticity of the digital object including that obtained from the Submission Information Package (SIP) during Ingest. As defined in the OAIS model, this authenticity evidence forms the Preservation Description Information (PDI) about the digital object and consists of various types of information including Reference, Context, Provenance, Fixity and Access Rights. The main underlying idea is to help to ensure that appropriate provenance is captured, for example if the data is transformed to a different format. Provenance is used to assess the Authenticity of a particular digital object.

4.1.6 Packaging Toolkit

The Packaging Toolkit is used (mainly during Ingest) to construct AIPs that will be stored using the Storage Service (see Section 4.1.7). The information collected in an AIP is aggregated either physically, or more likely, logically. In the latter case, the toolkit identifies within the AIP the location of the components so that they can be instantiated as a physical object for dissemination when requested by user. Additionally, the packaging toolkit needs to interact with the RepInfo toolkit to identify and obtain the RepInfo to accompany the digital object in an AIP.

4.1.7 Storage Services

This service provides an interface to the physical storage of digital objects. Using this interface ensures that all the information needed for the long term preservation of the data is identified (in an AIP) and can be moved from a repository to another when, for example the funding for the former ends. The interface can be implemented on top of existing storage systems so there should be no need to make major changes in existing repositories – it just adds the AIP capabilities. New storage systems could also be adopted, though this would not be without costs. For example in the last years, storage services have been progressively moving to web-based platforms in which the user sees a virtual archive that seamlessly takes care of all storage functions (data distribution, redundancy, refresh, etc.). Cloud storage is the technological basis for this service, which hides the physical storage complexity.

Therefore, in the SCIDIP-ES project, the aim is not to develop a new storage service for the ES data but to provide the data holders with "preservation-aware" storage service infrastructure based on existing storage technologies including cloud-based services.

4.1.8 Persistent Identifier Service

The ability to unambiguously and persistently locate and access digital objects is an important requirement of successful long-term digital preservation. In a digital preservation archive, the use of persistent identifiers (PIs) is ubiquitous including identifying AIPs in the storage as well as the RepInfo records in the RepInfo Registry. Assigning PIs to objects is usually the task of the **Data Management** component ([1]) of the archive.

In SCIDIP-ES, we aim to develop a simple persistent identifier service that interfaces to multiple existing Persistent Identification (PI) systems (e.g. DOI[8]) to obtain a unique identification code for the digital objects that are created within the system. It allows the interoperation of persistent identifiers used in different repositories and spreads the risk associated with an single PI system.

4.1.9 Orchestration Service

The Orchestration Service provides a brokerage service between existing data holders and their successors. Additionally, it also serves more generally as a knowledge broker. In particular it can exchange intelligence about events which might impact the long-term usability and/or access of data, e.g. changing technologies (support for new media and data formats), changing terminologies/knowledge of the DC and even changing ownership of data/ archive. Each of these kinds of changes may bear certain preservation risk concerning the data holdings in question. The Orchestration Service is intended to act as a collector of information about these kinds of events and broker the corrective actions necessary.

4.1.10 Finding Aid Toolkit

To support users' need to access and use data from many sources across many domains the infrastructure will provide a Finding Aid Toolkit to supplement the many existing domain search facilities. The development of this toolkit will aim to address, by utilising and harmonising related metadata and semantics (ontologies), the discovery of ES data that are not easily discoverable and accessible as they are heterogeneous in nature, (e.g. data coming from different sensors on different platforms such as satellites, aircraft, boats, balloons, buoys or masts, or located on the land), they are spread all over the world and originate from different applications.

This toolkit is not strictly related to data preservation process but it is a fundamental instrument to allow digital objects to be discovered by users and can play a fundamental role when it comes to data interoperability between different user communities.

4.1.11 Data Virtualisation Toolkit

The Data Virtualisation Toolkit allows the curators to inspect and describe the contents and structure of a digital object in a format independent manner creating the appropriate RepInfo. For example, in principle, using the toolkit, the contents of a

[8] Digital Object Identifier - http://www.doi.org/

NetCDF-based file could be viewed in a tabular format without needing a dedicated NetCDF viewer. In addition, the toolkit could also be used to help create (using the RepInfo toolkit) further RepInfo about any sub-components of the object as part of preservation planning and analysis. We envisage that this could also facilitate data access - i.e. consumers could use this type of RepInfo to bring together and analyse data from multiple sources without having to use multiple dedicated software systems. If full analysis capabilities are not available in this way the at least the consumer could inspect the actual content of a digital object before making the effort to obtain all the RepInfo needed to use it.

4.1.12 Process Virtualisation Toolkit

Process Virtualisation Toolkit is of fundamental importance in cases where digital objects need to be re-processed in the future to generate added value products. Thus, all information and/or ability to perform the digital object processing need to be preserved as well. The process virtualisation describes, in general terms, the various processes associated with the data by enabling an archivist or repository manager to identify the missing pieces of a given processing chain and apply corrective actions. For example, it may be necessary to re-compile the source code in order to run it in a different infrastructure ("create L-1C product from L-1B and port to new processing environment") as well as instantiating virtual host on-demand for processing.

4.1.13 Certification Toolkit

The Certification Toolkit will be a relatively simple tool for collecting evidence based on the ISO 16363[9] draft standard to submit for the ISO certification process. In addition, this tool will also be useful for self-audit and preparation for full, external audits.

4.2 Initial Prototypes and Validation

In the initial phases of SCIDIP-ES, we have developed basic prototypes of six of the services and toolkits - RepInfo Registry, Gap Identification Service, RepInfo Toolkit, Packaging Toolkit, Orchestration Manager and Data Virtualisation Toolkit. The development of these initial prototypes have been based on their original implementations by the CASPAR project, which also produced an extensive collection of evidence of their effectiveness in terms of preservation in several science disciplines (STFC and ESA repositories), cultural heritage (UNESCO world heritage[10]) and contemporary performing arts (INA, IRCAM, ULeeds and CIANT).

In SCIDIP-ES, these prototypes have undergone further evaluation by the projects partners, in particular the ones with ES data holdings with a view to understanding how the prototypes would be used in their archives. As a key outcome, this evaluation identifies the need for the services and toolkits to be more robust, scalable and simplified, where possible. These prototypes are publicly accessible for review[11].

4.3 Key Implementation Challenges and Future Work

As mentioned above, the majority of the services and toolkits in SCIDIP-ES were originally designed and implemented as proof-of-concept prototypes by the CASPAR project. In SCIDIP-ES, we aim to turn the CASPAR prototypes into production quality services, that is operational, scalable and robust as well as simplified (where possible) software products. In the process we will re-design the specifications based on the user cases and requirements defined in the project.

To address the scalability requirement of the RepInfo registry service, we aim to develop a network of RepInfo Registry services, section 4.1.1 to enable load distribution of requests for RepInfo between multiple registries acting as “Nodes” in the network. In order to avoid a single point of failure, all the registries will be essentially identical, apart from their holdings of RepInfo. There will be at least one registry, the Guarantor Node, in the network which we guarantee will be running even if all the others close down. The name(s) of the Guarantor node(s) will be propagated (e.g. via configuration in registry.representation.info property) so that new registries can register themselves with it.

The Gap Identification Service needs to improve the speed of query processing and providing support for transitive queries (a rule-based approach is investigated in [11]), while the Orchestration Service requires improved and more efficient support for the notification of preservation related events.

For the toolkits, such as Authenticity, Provenance and Integrity Toolkit and the Preservation Strategy Toolkit, we will aim to incorporate scalability in the underlying information models. Our analysis indicates that scalability of this toolkit could be achieved by creating a PNM record per group or collection of digital objects rather than per digital object in an archive. We are also exploring the feasibility of profiling the PREMIS metadata model [5] in the form of an OWL ontology to enable automation of creation and management of PNM records.

In addition, the development of the Persistent Identifier Interface Service will collaborate with the APARSEN project[12] that is developing an interoperable framework for persistent identifier services. We will also leverage the work done by the SHAMAN project, particularly the SHAMAN Preservation Context Model [6], for addressing the scalability and other related issues associate with the Process Virtualization and Emulation, and Packaging Toolkit. Similarly, we plan to build on the CASPAR Preservation Data Store [7] interface and the Kindura project's [8] approach to integrating traditional data system with Cloud-based technology in order to develop "preservation-aware" interfaces to any suitable existing storage systems. In effect, this would serve as the Storage Services needed for the SCIDIP-ES services and toolkits.

Besides the scalability and robustness issues, we have a number of other challenges. The infrastructure and toolkits must be usable in a number of existing systems. For instance, we plan to build the data virtualisation toolkit for ES data as a uniform front-end on a variety of existing data specific tools, including those creating standardised data descriptions using specifications such as EAST[13] and DFDL[14]. To verify this we will show effective integration in several different repositories, supporting their decision making and respecting their existing hardware, software, and governance and control systems. We must be able

[9] ISO 16363 - http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510
[10] UNESCO world heritage - http://whc.unesco.org/
[11] http://jenkins.scidip-es.eu/joomla/.
[12] The APARSEN Project - http://www.alliancepermanentaccess.org/index.php/aparsen/
[13] The Data Description Language EAST - A Tutorial - http://public.ccsds.org/publications/archive/645x0g1.pdf
[14] Data Format Description Language - http://www.ogf.org/dfdl/

to create RepInfo to enable broader use of the data. This will be verified as we enable ES users in different disciplines to find and use each other's unfamiliar data.

5. CONCLUSIONS

In this paper we have discussed the motivations and approach in the design of a preservation infrastructure, initially targeted at ES, but which has wider application across scientific disciplines. There are two aspects to this which we wish to highlight.

Firstly, that there needs to be a thorough preservation analysis to establish the context in which the preservation initiative takes place. For scientific data such as Earth Sciences this is an analysis beyond the digital objects themselves to consider both the dependencies on other entities that provide contextual information, which themselves have dependencies to form a network, and the requirements and assumptions of the designated community. From this analysis, an assessment of the costs and benefits of the preservation can be undertaken, taking into account the risks involved in preservation. In SCIDIP-ES we are demonstrating the value of this approach in practise in the implementation of the use cases, via the use of Preservation Network Models.

An important outstanding aspect of this approach is the establishment of the value of preservation for the archives involved. This is necessarily difficult to assess; it is particularly difficult to give the value of the use of data to support future scientific advances and subsequent impact on society. However, a framework for estimating the value proposition is nevertheless required to justify the additional effort of preservation, which does tend to be front-loaded. For archives such as those involved in SCIDIP-ES, which are from publically supported science, the framework should extend beyond the purely commercial to cover research and ultimately societal benefits. An ongoing work item within SCIDIP-ES is exploring such a framework.

Secondly, in order to support effective preservation, an infrastructure with a number of services needs to be provided to support the stages of the OAIS functional model. In this paper, we have outlined the services identified within SCIDIP-ES and discussed how they might interact to support a preservation scenario. To be sustainable, these services must be of production quality. We have discussed their scalability, in both size and heterogeneity. But to be production quality these services also need to be robust in the presence of failure, secure to maintain the integrity of the data, and maintain accessibility as the environment changes. Thus to be sustainable, these tools themselves need to adhere to a strong preservation discipline.

SCIDIP-ES has completed its initial analysis and design, as reported in this paper. In the next phases of the project, this sustainable infrastructure will be realised, deployed, and evaluated on the ES use cases.

Finally, it should be noted that the SCIDIP-ES preservation services and toolkits are designed for much wider application than the Earth Science use cases considered in this paper. For instance, we envisage that the SCIDIP-ES infrastructure, when developed, will have the potential to aid long-term preservation of data exposed through the large-scale Spatial Data Infrastructures (SDIs) in Europe, such as the INSPIRE SDI[15], which currently do not address preservation [9].

[15] INSPIRE - http://inspire.jrc.ec.europa.eu/

6. ACKNOWLEDGMENTS

The work presented in this paper is funded by the Seventh Framework Programme (FP7) of the European Commission (EC) under the Grant Agreement 283401.

7. REFERENCES

[1] CCSDS. 2002. Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data Systems Standard, *Consultative Committee for Space Data Systems (CCSDS) Blue Book*, 2002, or later URL: http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Y. Marketakis and Yannis Tzitzikas 2009: Dependency management for digital preservation using semantic web technologies. Int. J. on Digital Libraries 10(4): 159-177 (2009)

[3] Conway, E., Dunckley, MJ., Giaretta, D. and Mcilwrath, B. 2009. Preservation Network Models: Creating Stable Networks of Information to Ensure the Long Term use of Scientific Data, *Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data,* del Castillo, Madrid, Spain, 01-03 Dec 2009. URL: http://epubs.cclrc.ac.uk/bitstream/4314/PV09_Conway_PNM.pdf

[4] PREMIS. 2008. PREMIS Data Dictionary for Preservation Metadata, version 2.0, *PREMIS Editorial Committee*. URL: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[5] Brocks, H., Kranstedt, A., Jäschke, G., and Hemmje, M. 2010. Modeling Context for Digital Preservation. *In E. Szczerbicki & N. T. Nguyen (Eds.), Smart Information and Knowledge Management* (Vol. 260, pp. 197–226). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04584-4_9

[6] CASPAR Consortium. 2008. D2201: Preservation Data Store Interface, *CASPAR Consortium*, 2008. URL: http://www.casparpreserves.eu/Members/cclrc/Deliverables/preservation-datastore-interface/at_download/file.pdf

[7] Waddington, S., Hedges, M., Knight, G., Zhang, J., Jensen, J. and Downing. R. Kindura. 2012. Hybrid Cloud Repository, *Presentation*, 2012. URL: http://www.jisc.ac.uk/media/documents/events/2012/03/waddington_kindura.pdf

[8] Shaon, A., Naumann K., Kirstein M., Rönsdorf C., Mason P., Bos M., Gerber U., Woolf A. and G Samuelsson G. 2011. Long-term sustainability of spatial data infrastructures: a metadata framework and principles of geo-archiving, *Proc. 8th International Conference on Preservation of Digital Objects*, Singapore, 01-04 Nov 2011. URL: http://epubs.stfc.ac.uk/bitstream/7195/GeoPres_IPRES_CR.pdf

[9] PARSE.Insight. 2010. Case Study Report, *PARSE.Insight Public Report*, 2010. URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf

[10] Y. Tzitzikas, Y. Marketakis and Y. Kargakis, 2012, Conversion and Emulation-aware Dependency Reasoning for Curation Services, iPres 2012