SCIDIP-ES: a Sustainable Data Preservation Infrastructure to Support OAIS Conformant Archives

Shirley Crompton[1], David Giaretta[1,2], Brian Matthews[1], Holger Brocks[3], Felix Engel[4],
Arif Shoan[1,] Fulvio Marelli[5]

[1]Scientific Information Group, Science and Technology Facilities Council, UK
{shirley.crompton, david.giaretta, brian.matthews,
arif.shoan}@stfc.ac.uk
[2]Alliance for Permanent Access, UK
director@alliancepermanentaccess.org
[3]InConTec GmbH, Germany
holger.brocks@incontec.de
[4]FTK, Germany
fengel@ftk.de
[5]European Space Agency, Italy
Fulvio.marelli@esa.int

Abstract. The SCIDIP-ES project provides an e-infrastructure to support long-term preservation and use of the knowledge encoded in scientific data. The infrastructure offers a set of generic, sustainable services and toolkits based on the CASPAR prototypes to support efficient preservation planning and management along with usability and access needs. The SCIDIP-ES services are specifically designed to augment existing repositories as well as to facilitate the development of new ones, so that they can be properly OAIS conformant as described in section 1.4 of the OAIS Reference Model. In particular, the infrastructure enables repositories to support the OAIS Information Model and to fulfil several of the mandatory responsibilities which are required for conformance yet which are rarely discussed.

Keywords: digital preservation, OAIS, e-infrastructure, earth science, services.

1 Introduction

The EU regards long term preservation of scientific data a public interest in view of the societal and economic gains from improved access to and continuous exploitation of reliable research output [1]. SCIence Data Infrastructure for Preservation – Earth Science (SCIDIP-ES) [2] supports this vision by offering generic, sustainable services and toolkits for efficient preservation planning and management along with usability and access needs. By helping to build a broad Earth Science (ES) user community of significant mass, sharing and using each other's data over the long term, implementing the software in a simple and robust fashion, and linking to the

Alliance for Permanent Access (APA) [3] beyond the project, we aim to guarantee sustainability of these services. Besides being driven by requirements from the Reference Model For An Open Archival Information System (OAIS) [4], the services offered are those which have been shown by CASPAR [5] as being effective for preservation and by PARSE.Insight [6] as being widely recognized as needed by diverse communities. To ensure the SCIDIP-ES infrastructure will indeed be used (i.e., we wish to avoid a "build it and they will come" approach), the detailed requirement specifications and customisations of these services and toolkits have been defined using various exemplars from ES and other domains.

Science data infrastructures, e.g., within the Earth Science (ES) domain, are inherently heterogeneous and fragmented due to the proliferation of completely different data capture instruments and data formats used. In addition, most of these existing science data infrastructures are not OAIS conformant, i.e. not preservation-aware. Shaon *et al.* [7], for instance, have reported on the main preservation challenges and barriers that science data infrastructures need to overcome. SCIDIP-ES can help data archives to become preservation-aware through adopting its OAIS-conformant services and toolkits. In this paper, we describe the design and functional specification of the SCIDIP-ES infrastructure and, in specific, how we built-in support for the OAIS standards. We begin with a brief characterization of the state of OAIS conformance within digital preservation (Section 2), followed by a review of the mandatory responsibilities required for conformance (Section 3). Section 4 describes the three high level use cases developed using results from analyses of ES exemplars and responses to a requirement survey [8] conducted by the project. The analyses were aligned with the OAIS Reference Model, focusing on identifying the roles, responsibilities and functionalities in archival information preservation. The resultant use cases and their requirements were then used to guide the functional definition of SCIDIP-ES services and toolkits. Their intended application is illustrated according to these use cases (Section 5). Finally, we conclude with a report on our development plan and the evaluation approach for testing the effectiveness of the infrastructure.

2 The State of OAIS Conformance within Digital Preservation

The field of long-term digital preservation has witnessed marked progress over the last few years. This has predominantly included newer preservation tools and systems being developed at both academic and commercial spheres to address the technical challenges of digital preservation. However, most adhere to the mantra "emulate or migrate", ignoring the more general aspects of the OAIS Information Model, in particular the use of Representation Information (RepInfo) to help encapsulate knowledge and describe digitally encoded information. In this respect, semantic RepInfo is sadly neglected in terms of preservation.

The currently on-going SCAPE project [9] is building a series of preservation tools and services to address the scalability challenges of various preservation operations including ingest, characterization and migration while the ENSURE project [10] focuses on emulation in virtual machines. More recently, there has also been an increased focus on a few conceptual areas of work, particularly in terms of preservation policy models (e.g. SCAPE), and, to some extent, preservation cost estimation models (e.g. ENSURE). Unsurprisingly, the majority of these endeavors are inspired or influenced and, in some cases, even underpinned by the OAIS Reference Model. It should be noted that the Model is a conceptual model of the functions and responsibilities of an archive viewed as an organization or other entity charged with the long term preservation of digital contents. Although it articulates the preservation processes, roles and responsibilities in an OAIS, it does not actually describe how to achieve this. Thus, an issue with the trend of preservation tools and services (both established and emerging) adopting the Model is that developers have different and, often, partial interpretations of the Model. For example, both the Safety Deposit Box (SDB [11], a commercial preservation system developed by Tessella [12]) and Roda [13] (an Open Source and research-based preservation repository developed by the Portuguese National Archives) claim to be OAIS-compliant. However, neither seems to support a RepInfo Network (see Section 5.1) linked to a defined Designated Community (DC) - an approach adopted by SCIDIP-ES (see Preservation Archive Creation). In the following sections, we present how SCIDIP-ES addresses OAIS requirements.

3 Responsibilities of an OAIS Conformant Archive

OAIS defines a set of mandatory responsibilities (OAIS Section 1.4) that digital preservation archives should fulfill in order to be conformant. These include:

- A conforming OAIS Archive implementation shall support the model of information described in 2.2 *[of OAIS]*. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.
- A conforming OAIS Archive shall fulfill the responsibilities listed in section 3.1 *[of OAIS]*.

For the latter we focus on:

(a) Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.

(b) Ensure that the information to be preserved is *Independently Understandable* to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.

(c) Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.

(d) Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.

The SCIDIP-ES e-infrastructure provides 13 generic services and tools (see Section 5) that facilitate fulfilling the above requirements. The remaining two OAIS responsibilities concern the operational policy of a preservation archive and are therefore outside SCIDIP-ES' scope:

(e) Negotiate for and accept appropriate information from information Producers.
(f) Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.

The SCIDIP-ES infrastructure is designed for use by any organization involved with long-term digital preservation. However, our primary focus is to showcase its use in the context of ES organizations working with non-ES ones concerned with data preservation to confirm SCIDIP-ES' broad effectiveness in helping to improve, and reduce the cost of, the way in which they preserve their ES data holdings. In the next section, we describe the SCIDIP-ES use cases which encapsulate the OAIS responsibilities 3(a-d) identified above and which have been contextualized with reference to ES community-specific requirements.

4 The SCIDIP-ES Use Cases

To ensure conformance, the OAIS Reference Model is consistently used in the design and functional specifications of the SCIDIP-ES services and toolkits. For instance, we aligned the three SCIDIP-ES high level use cases to the Reference Model to capture functional requirements in the long-term preservation of ES data. These use cases reflect the three distinct phases in the OAIS lifecycle model of data preservation:

- **Preservation Archive Creation** - to support the processes for setting up an archive or a new collection of data within an existing archive. These include identifying what kind of information need to be preserved to ensure the long-term usability of the ES data by the target DC; determine the user roles and preservation objectives; define a cost-effective preservation strategy and the correct procedures needed to implement the archive. For existing archival systems, this would also need to address the efficient integration of preservation processes within the existing system architecture.
- **Archived Data Access** - this relates to the access and exploitation of the preserved data. The functionalities required are search, discovery and retrieval of associated information in the archive, to allow data consumers to access, interpret and use the preserved data efficiently, correctly, and ideally within their familiar tools.
- **Archive Change/Evolution** - to protect the preserved data against changes which could range from the DC, technology, policies to funding issues etc. In the case of a changing DC, an archive may choose to update or augment the RepInfo associat-

ed with the preserved data to ensure that it is still intelligible and usable by the new DC. In the extreme case of an archive ceasing operation, it would need to identify and prepare to handover its holding to a suitable successor. Whatever the nature of the challenge, an archive should have the capability to plan responses to changes in a safe, cost-effective and sustainable manner.

5 Mapping SCIDIP-ES Use Cases to Infrastructure Services and Toolkits

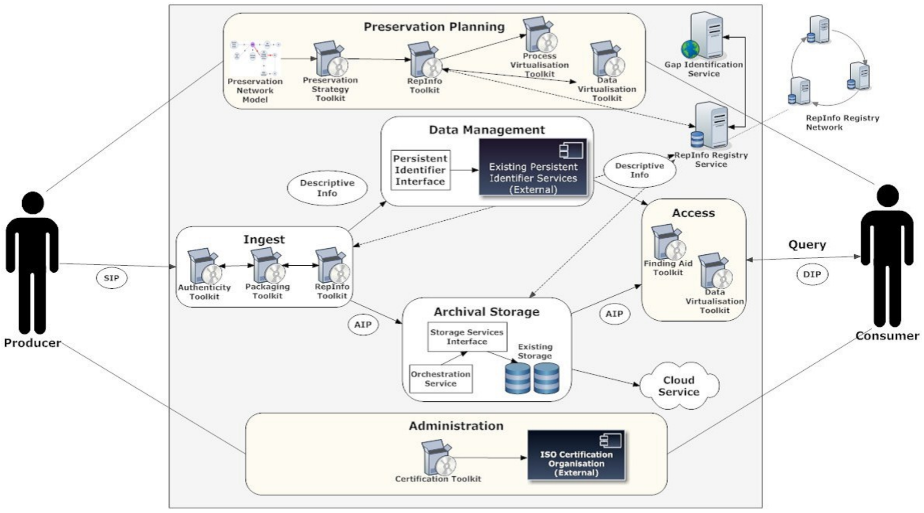

Fig. 1. Overview of SCIDIP-ES services and toolkits.

Figure 1 gives an overview of the SCIDIP-ES services and toolkits aligned to the familiar OAIS Functional Model. Although ES is used as the pathfinder for building the infrastructure, it should be stressed that the component services and tools are designed to support the full preservation life-cycle irrespective of the data domain as defined in the uses cases above. In the next sections we motivate the key services and toolkits through the use cases and by reference to the needs of OAIS conformance.

5.1 Preservation Archive Creation

To conform to OAIS an archive must implement the OAIS Information Model and also fulfill, amongst others, the mandatory requirements about independent usability by its identified DC [see 3(a,b,d)], and the ability to hand over its digital holdings to a successor [see 3(c)]. In SCIDIP-ES, we see the ability to construct complete Archival Information Packages (AIPs) as key to the above OAIS requirements. We must of course bear in mind that an archive is unlikely to completely change the way in which it stores its data holdings and the related Preservation Description Information (PDI), both of which, experience shows, archives do possess. However, there is commonly no definition of the DC for which the data is being preserved nor of adequate RepInfo to support its usage. While this is marginally acceptable for rendered holdings, this is unacceptable for long term data preservation because the latter requires that data be process-able and can be combined with other, perhaps newer, data. Therefore we provide toolkits to create AIPs using functionalities provided by the **Packaging Toolkit** and **Storage Service.** For an existing archive, elements in the AIP would reference locations within the archive, and to RepInfo.

The amount of RepInfo required is measured against the perceived skills, resources and knowledge base available to the target DC. An archive may wish to broaden exploitation of its data holding by providing additional RepInfo for a wider group of users with different knowledge and resource bases. Given the potential diversity and quantity of RepInfo involved, an archive cannot by itself be expected to capture and manage all the RepInfo that it might require. To support these key requirements and help share the burden and efforts for preserving data long term, SCIDIP-ES provides the **RepInfo Registry Service** which would be used to query, retrieve and manage RepInfo required by a group of preservation archives. Note that we fully expect that repositories will locally cache the RepInfo they need. To facilitate the use of RepInfo, the Registry will contain a special type of RepInfo called RepInfo Labels (RILs) which are pointers to multiple RepInfo objects. We note that a RepInfo Registry must itself be an archive, and so, logically, the actual RepInfo data objects would be preserved as OAIS AIPs which include PDI and its own RepInfo objects to facilitate interpretation. The latter construct gives rise to RepInfo Networks (RINs) and the Registry would also enable users to navigate a network of RepInfo objects to explore the knowledge represented. In SCIDIP-ES, an RIN represents the chosen solution for fulfilling a specific preservation objective (see next section).

The **Gap Identification Service (GIS)** is defined to help assess if a data consumer with a particular knowledge profile can 'understand' the preserved digital objects by identifying "gaps" in the corresponding RIN stored in the Registry [14]. Preservation planning and the creation of RepInfo are also ubiquitous activities for this use case and these are discussed in more details in the Archive Change Evolution section (5.3).

To assess if the defined archive is OAIS conformant and to identify potential areas for improvement, archivists may use the **Certification Toolkit** which implements the ISO16363 standards for Trustworthy Digital Repositories Audit and Certification [15] to perform a self-audit.

5.2 Archived Data Access

To meet users' need to discover, access and use data from different sources [see OAIS responsibilities described in 3(b)] across multiple domains, SCIDIP-ES defines a **Finding Aid Toolkit (FAT)** to support the many existing domain search facilities by building on an archive's current search and find capabilities. As highlighted in 3(d), preserved information should be disseminated with sufficient provenance information to provide quality assurance. SCIDIP-ES offers an implementation of the **Authenticity Toolkit (AT)** for associating provenance evidence and other PDI-related records on ingest. As in the previous use case, the Registry and GIS services could ensure that adequate RepInfo is available given the user's registered DC knowledge profile and the requested data.

It is envisaged that the **Data Virtualisation Toolkit** (DVT, also see next section) may also be used in conjunction with specific types of RepInfo to facilitate data access. The toolkit would allow users to inspect the contents and structure of a digital data object using the associated semantic and structural RepInfo, eg. viewing a NetCDF-based [16] file in tabulated format without a dedicated NetCDF viewer. In this way, users are able to bring together and analyze data from multiple sources without having to use multiple dedicated software systems [see OAIS responsibilities described in 3(b)]; or preview data content of the preserved digital object before making the effort to obtain all the RepInfo needed to use the data.

5.3 Archive Change/Evolution

Changes are inevitable given the timescale involved in preserving data long-term and an OAIS conformant archive must ensure that information is preserved against all reasonable contingencies [see 3(c)]. To monitor changes to technology or DC that might affect the long-term accessibility and usability of the preserved data, SCIDIP-ES provides the **Orchestration Service (OS)** to act as a knowledge broker which preservation archives may subscribe to receive notifications on specific topics of interest. The source of such information must come from experts in the various fields.

Those experts may also be asked to create additional RepInfo and so we provide the **RepInfo Toolkit (RIT)**, a user-friendly GUI that is actually a collection of tools to facilitate the creation of RepInfo objects and interactions with the Registry. Some sub-components of this toolkit, e.g. the Data and **Process Virtualisation Toolkit**s (**PVT**), are aimed at describing the data in more "virtualized" terms to help integrate the data into other software or interoperate data from multiple sources.

In terms of preservation, there are a number of basic strategies for preserving digitally encoded information. Besides using RepInfo (which includes emulation software), one could migrate (Transform in OAIS terminology) the data into different formats. CASPAR developed the Preservation Network Model (PNM) [17] which helps an archivist to plan and evaluate different preservation strategies, balancing factors such as costs against efficacy and risk against tolerance for given specific preservation objectives. The **Preservation Strategy Toolkit** (**PST**) provides an intuitive simple to use GUI for archivists to build PNMs as network diagrams, to capture the relationships between the PNM objects together with their attributes such as location, preservation objectives, risk, cost, tolerance and quality assurance, etc.

As an illustration of the notion of PNM and PST, a digital object might depend on multiple information objects (RepInfo) for a particular function or operation that could be performed on or using that object. The relationships between the objects could be either composite or alternate, where the former signifies a dependency on a combination of digital objects, while the latter represents optional dependencies. For example, to use a piece of scientific data, like combining it with newly acquired data, there are two basic approaches:

- Transform the data into a format compatible with the analysis tools used for the new data – this has associated computational and storage costs, and potential loss of information. OR
- Describe both sets of data in such a way that software can access and combine information from both. For example, describing the bit location and encoding of temperature measurements from the same geographical location for each dataset.

Whichever alternatives is used, both will require that the user is provided with Semantic Representation such as the units of the temperature measurements, whether they refer to surface or sub-surface temperatures and so on. Thus, recording these types of relationships is vital as they play an important role in preservation strategy formulation and evaluation in an archive. Both approaches have its own costs and risks, and these may be different for different archives for the same digital holdings. PST supports evaluation of alternative solutions within a PNM by the user defined risk and or cost profiles. PST could also be used to monitor the stability of an implemented PNM solution (in the form of an RIN in the Registry). For example, technology changes may render a RepInfo obsolete leading to a breakdown in the RIN. PST could be used to re-evaluate the original PNM in view of the new information to identify a new solution which may simply involve using an alternative relationship already defined in the PNM.

Long-term data preservation requires long-term commitment. It may come the point when an archive needs to handover ownership of its data due to a change in policy or funding. To facilitate the changeover process, OS could also play the role of a generic knowledge broker that could be used to identify a suitable successor. In SCIDIP-ES, both RepInfo and data objects are preserved as OAIS AIPs. This approach ensures that the tacit dependencies and knowledge are captured in a standard format [see OAIS responsibilities described in 3(c)] to guarantee the continuous accessibility and usability of the preserved data.

6 Summary and Future Work

In this paper we described the objectives of the SCIDIP-ES project which include delivering a sustainable OAIS conformant infrastructure for making data archives preservation-aware. In particular, we discussed in details how the SCIDIP-ES services and toolkits meet mandatory OAIS responsibilities through illustrating their usage within the data preservation lifecycle as represented by the three high level SCIDIP-ES use cases. SCIDIP-ES is now in its final year and has already delivered two releases of the software. The focus of the remainder of the project is to conduct intensive testing of the services and toolkits to collate feedback and refinement requirements to feedback into the last iteration of the product scheduled for release in February 2014. To ensure that the services and toolkits are useful for long-term knowledge preservation as defined by the OAIS Reference Model and can be handover as well as sustained, we are developing customized implementation with different configurations of the services and toolkits to integrate with selected ES partners' data archives in different countries, including ESA [18], DLR [19], BGS [20] and BADC

[21], etc. These custom implementations will serve as test beds for us to test the usability of the SCIDIP-ES software in different contexts with reference to functional and quality requirements. For example, ESA wishes to run their own Registry and a more complete set of SCIDIP-ES services and toolkits while BGS and BADC are willing to share the use of an external Registry and the efforts of maintaining commonly used RepInfo objects. Finally, we re-iterate that the SCIDIP-ES infrastructure, though initially targeted at ES, is intended to have wider application across scientific disciplines. To achieve this, we designed the services and toolkits to implement the OAIS Information Model and provide generic functionalities to support the relevant mandatory OAIS responsibilities – which are domain agnostics

Acknowledgement

SCIDIP-ES is funded by the Seventh Framework Program (FP7) of the European Commission (EC) under the Grant Agreement 283401.

References

1. EU: Commission Recommendation of 17 July 2012 on access to and preservation of scientific information. Official Journal of the European Union (2012), http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32012H0417:EN:NOT
2. Science Data Infrastructure for Preservation – Earth Science (SCIDIP-ES) Project, http://www.scidip-es.eu/
3. Alliance for Permanence Access, http://www.alliancepermanentaccess.org/
4. CCSDS: Reference Model for an Open Archival Information System (OAIS). Recommendations for Space Data Systems Standard, Consultative Committee for Space Data Systems (CCSDS) Magenta Book (2012), http://public.ccsds.org/publications/archive/650x0m2.pdf
5. Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval (CASPAR) Project: http://www.casparpreserves.eu/
6. PARSE.Insight Project: Case Study Report, PARSE.Insight Public Report (2010), http://www.parse.insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf
7. Shaon, S., Giaretta, D., Crompton, S., Conway, E., Matthews, B., Yu, J., Marelli, F., Di Giammatteo, U., Marketakis, Y., Tzitzikas, Y., Guarino, R., Brocks, H. & Engel, F.: Towards a Long-term Preservation Infrastructure for Earth Science Data. The Ninth Annual Conference on Digital Preservation (IPres2012), Toronto, Canada (2012)
8. SCIDIP-ES: Deliverable 12.1 Requirement Specification and Gap Analysis Report, http://www.scidip-es.eu/assets/Deliverables/SCIDIP-ES-DEL-WP12-D12-1.pdf
9. SCAlabe Preservation Environments Project: http://www.scape-project.eu/
10. ENSURE Project: http://ensure-fp7-plone.fe.up.pt/site
11. Safety Deposit Box: http://www.digital-preservation.com/solution/safety-deposit-box/
12. Tessella: http://www.digital-preservation.com
13. Roda: http://roda-community.org/what-is-roda/
14. Marketakis, Y., Tzitzikas, Y.: Dependency Management for Digital Preservation using Semantic Web Technologies, International Journal on Digital Libraries, 10(4), (2009).

15. ISO 16363:2012: Audit and Certification of Trustworthy Digital Repositories (2012), http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510
16. UniData: Network Common Data Form (NetCDF), http://www.unidata.ucar.edu/software/netcdf/
17. Conway, E., Dunckley, M.J., Giaretta, D., McIlwrath, B.: Preservation Network Models: Creating Stable Network of Information to Ensure Long Term Use of Scientific Data. Proceedings Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data, del Castillo, Madrid, Span (2009)
18. European Space Agency (ESA): http://www.esa.int/ESA
19. Deutsches Zentrum Fuer Luft – und Raumfahrt EV (DLR): http://www.dlr.de/dlr/en/desktopdefault.aspx/tabid-10002/
20. British Geological Survey (BGS): http://www.bgs.ac.uk/
21. British Atmospheric Data Centre (BADC): http://badc.nerc.ac.uk/home/index.html