Core Strategic Vision

1 Introduction

1.1 Purpose of this document

This document has been prepared for EU Concertation Meeting 24th Nov 2008 and as a contribution to the Roadmap for the Alliance for Permanent Access and the PARSE.Insight project.

1.2 Preservation Infrastructure: integration with and differentiation from other infrastructures

Preservation Infrastructure is taken here to mean those things, technical, organization and financial which are usable across communities to help in the preservation of digital holdings. The focus of this Roadmap is largely at the technical level but the other aspects

are also addressed briefly. Preservation is meant in the OAIS[1] sense of maintaining the usability and understandability of a digital object[2].

Community-specific infrastructures, adapted to the needs of organizations within specific communities, are possible but should use and complement the services of the more general infrastructure.

This preservation infrastructure must integrate with the computation and data GRID-type infrastructure and provides analogous functionality in the sense of providing the linkage between islands of resources, as shown in Figure 1.

The infrastructure components provide the linkage between islands of capabilities just as the network infrastructure (e.g. GEANT) links national networks and compute infrastructures (e.g. EGEE) link islands of compute and storage resource. The Preservation infrastructure links islands of capabilities separated by time. This is a one way communication from present to future and in this case there are a number of threats which hinder the correct transmission of digitally encode information.

It should be noted that there is a fundamental difference between the preservation infrastructure components and some or all of the rest of the infrastructure. This arises because there is a requirement, by definition, of a long-term commitment. By contrast middleware GRID systems quite naturally have shown a rapid turnover and lack of long-term commitment to any individual system.

1.3 Terminology

Unless otherwise stated the terminology used comes from OAIS.

[1] the Reference Model for an Open Archival Information System, ISO 14721
[2] for a designated community

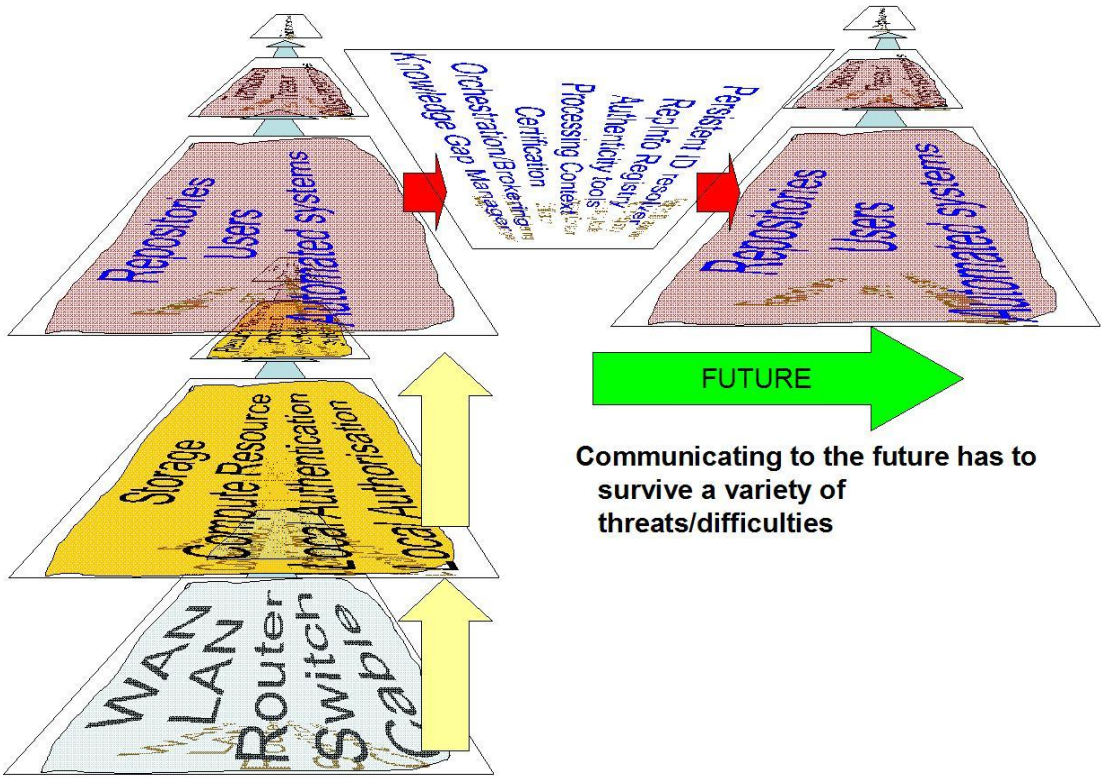

Figure 1 Infrastructure - including preservation

2 Demand for a Preservation Infrastructure

An associated paper summarizes the surveys which have been undertaken by PARSE.Insight and members of the Alliance for Permanent Access. These surveys show a substantial demand for a preservation infrastructure which is consistent across nations, continents and over a remarkably wide range of disciplines.

There has been time for only an initial analysis of the results. The results of most immediate interest revolve around a collection of "threats" to digital preservation which are based on prior analyses of the domain. It is worth noting that similar lists can be found in most project proposals related to digital preservation, e.g. compare the project descriptions of CASPAR, Planets, SHAMAN, etc.

The major threats are as follows:

1. Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved
2. Non-maintainability of essential hardware, software or support environment may make the information inaccessible
3. The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity
4. Access and use restrictions may not be respected in the future
5. Loss of ability to identify the location of data

6. The current custodian of the data, whether an organization or project, may cease to exist at some point in the future
7. The ones we trust to look after the digital holdings may let us down

The surveys show that between 50% and 70% of responses indicate that all the threats are recognized as either "Important" or "Very Important", with a majority supporting the need for an international preservation infrastructure.

The surveys have clearly not been conducted in a well sampled, rigorous way and there will naturally be a number of concerns about the validity of the results presented here. We have therefore addressed two pressing concerns, namely (1) that the survey results may be skewed by self-selection of the responders and (2) the threats may be either unsupported or else incomplete.

For the first of these we have shown that there is a surprising consistency of results when compared across different countries, continents and disciplines and organization types. Admittedly this is not a qualitative argument but nevertheless one we find very encouraging. In addition we have conducted a complementary telephone survey of non-responders to obtain some indication of whether their failure to respond indicates a major underrepresentation of the view that there is no demand for infrastructure.

To address the second concern we have analyzed the free text responses from individuals to questions about reasons for loss of data and we find no new threats but significant numbers of examples of each threat apart from one. The exception is threat number 4 above, namely that connected with rights management where it appears that the wording should have been "Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future" and we use this phrasing below

3 Requirements for a Preservation Infrastructure

We base the requirements for the preservation infrastructure on a broad analysis of the threats and an initial set of solutions.

Threat	Requirements for solutions
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	Ability to create and maintain adequate Representation Information
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	Ability to share information about the availability of hardware and software and their replacements/substitutes
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	Ability to bring together evidence from diverse sources about the Authenticity of a digital object
Access and use restrictions may	Ability to deal with Digital Rights correctly in a

make it difficult to reuse data, or alternatively may not be respected in future	changing and evolving environment
Loss of ability to identify the location of data	An ID resolver which is really persistent
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation
The ones we trust to look after the digital holdings may let us down	Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term

4 Technical Preservation concepts and components

Each of the solutions is analysed next in way analogous to the e-IRG Roadmap, with which this strategic vision should (eventually) be compatible. For each solution there is a need to review the existing digital preservation projects, review the proposals and identify open issues.

4.1 Create and maintain Representation Information

The information needed to understand and use a digital object is termed, in OAIS, "Representation Information". This is a catch-all term which includes information about a digital object's format, semantics, software, algorithms, processes and indeed anything else needed.

Next steps:

- Representation Information Registry holding copies of Representation Information of all types which can be shared and enhanced by contributions from many people.
- Knowledge Gap Manager which provides a semi-automated way of identifying where additional Representation Information needs to be created, based on information collected by the Orchestrator/Broker
- Processing Context which helps to maintain information about the processing history of a dataset

Final destination

- A set of services, supported over the long term, which make it easier to maintain adequate Representation Information, particularly after active work on the dataset has ceased or slowed.

Relevant projects, policies, organisations, activities:

- CASPAR, Planets, OAIS

4.2 Sharing of information about the availability of hardware and software and their replacements/substitutes

Next steps:

- Development and sharing of information about emulation and migration strategies
- Development of orchestrator/broker to share available substitutes
- Acts as (1) a clearing house for demands for Representation Information, (2) for collecting information about changes in availability of hardware, software, environment and changes in the knowledge bases of Designated Communities and, (3) to broker agreements about datasets between the current custodian, which is unable to continue in this role, and an appropriate successor.

Final destination

- A set of services which make it easier to exchange information about obsolescence of hardware and software and techniques for overcoming these.

Relevant projects, policies, organisations, activities:

- CASPAR, KEEP (with regard to emulation)

4.3 Ability to bring together evidence from diverse sources about the Authenticity of a digital object

Authenticity is not a Boolean concept. It is in general not possible to state that an object is authentic. Instead one can provide evidence on which a judgement may be made about the degree to which a person (or system) may regard an object as what it is purported to be. This evidence will be technical, for example details of what has happened to the object (Provenance) as well as social, for example does one trust the person who was in charge of the system under which the object has been held.

In general the provenance information associated with various objects will be encoded according to one of a multitude of different system e.g. CIDOC-CRM, OPM. There is at minimum a need to be able to interpret and present provenance evidence in a uniform way so that users can make an informed judgment about the degree of belief that a data object is what it is claimed to be. These tools would also facilitate the collection of appropriate evidence.

Next steps:

- Develop an authenticity formalism
- Develop international standards and common policies on authenticity and provenance.
- Creation of tools to capture evidence relevant to authenticity
- Tools to map provenance to authenticity tools

Final destination

- A set of standards and tools through which a user in the future can be provided with evidence on which he/she may judge the degree of Authenticity which may be attributed to a digital object.

Relevant projects, policies, organisations, activities:

- CASPAR, SHAMAN

4.4 Ability to deal with Digital Rights correctly in a changing and evolving environment

Allow the digital rights associated with an object to be presented in a consistent way, taking into account the changes in legislation. An associated problem is the circumstance in which the licence to access the object (or without which the required software is unusable) expires and the originating company no longer exists.

Next steps:

- Share information on how constraints, which DRM systems possibly impose on preservation planning and preservation actions, can be handled under different and changing legal systems
- Develop a dark archive for holding tools to generate licences, which would only be used if and when the commercial supplier is unable to provide this capability

Final destination

- Registry of/Clearinghouse for rights information and dark archive of licensing tools

Relevant policies, organisations, activities:

- CASPAR, ARROW (Accessible registries of rights information and orphan works towards Europeana)

4.5 An ID resolver which is really persistent

There is no shortage of things which are claimed to be Persistent Identifier systems. The issues associated with these are the scalability of the solutions and the longevity of the underlying organisational structure. A name resolving system whose persistence is guaranteed by an international, government based organisation is needed. This could build on one or more existing name resolving systems, strengthening the organisational structures underpinning the resolver.

Next steps:

- Review the existing persistent identifier systems and their technical, organisation and social underpinnings with respect to longevity and scalability
- Investigate potential organisational underpinnings and the links to, for example, the EU or USA.

Final destination

- An identifier system for digital objects which has adequate organisational, financial and social backing for the very long term which can be used with confidence

Relevant projects policies, organisations, activities:

- DOI, DNS, CASPAR, URN, nestor catalogue of criteria for trusted PI-systems, XRI

4.6 Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation

Projects and organisations can and do run out of funding for preserving digital holdings, for example projects from Earth Observation (EO) projects are often only funded for 10 years after the closure of the satellite from which the data is derived. There are in the EO case some more or less formal mechanisms for finding a host who could take over responsibility. A brokering/orchestration system is needed to formalise the finding of new hosts.

However even if agreement is reached there is the issue of collecting all the information related to a set of digital objects held, perhaps in a variety of systems, by the original host, and transferring this to the new host, itself with a variety of systems.

OAIS defines in very general terms an Archival Information Package which (logically) contains all the information needed for the long term preservation of a digital object. In addition to the Brokering/Orchestration mentioned above we need to be able to create the AIP so that these can be handed over to the new host.

Next steps:

- Create tools for collecting and (logically) packaging information into AIPs using information from a variety of underlying information systems
- Investigate the options for mapping systems from one major system to another.

Final destination

- A system which will allow organisations which are no longer able to fund the preservation of a particular dataset is able to find an organisation willing and able to take over the responsibility. The ultimate fallback could be the Storage Facility (see section 4.8.1.1)

Relevant projects, policies, organisations, activities:

CASPAR, SHAMAN, OAIS

4.7 Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term

Although one can guarantee anything into the indefinite future there, for more than a decade, been a demand for an international process for accreditation, auditing and certification of digital repositories, based on an ISO standard.

Next steps:

- Support the development of a set of ISO standards about digital repository audit and certification
- Help set up the organisation and processes to provide accreditation and certification services

Final destination

- An internationally recognised accreditation, audit and certification process with a well defined and long-lived support organisation, with appropriate tools and best practice guides.

Relevant projects, policies, organisations, activities:

- Repository Audit and Certification Working Group (http://wiki.digitalrepositoryauditandcertification.org), DRAMBORA, OAIS

4.8 Possible additional technical components

4.8.1 From e-IRG

4.8.1.1 Storage Facility

Provision of a network of distributed shared facilities will reduce overall costs as it takes away the need for inefficient local redundancy. The concentration of buying power and maintenance will also lower cost and increase quality, while having an installed base ready for use any time lowers deployment time. Grids are able to deal with sudden popularity of data, using the swarming effect (the consumer of data becomes part of the source). In short, it will allow for advanced data recovery faster than in any other scenario and at the lowest price possible – providing efficiency, flexibility, security, availability and scalability. With the networks and grid technologies in place to provide the interconnectivity and load balancing features, shared storage facilities are a key component in the grid equation.

Next steps:

- Design an optimal safe storage topology and determine a storage development roadmap.
- Link large distributed storage facilities able to replicate and serve grid data as a test bed.
- Find long term financial support for distributed European Storage Facilities.

Final destination

- A European Grid storage facility that is secure, distributed and extremely fast. This high capacity storage facility is at any given point in time capable of mirroring and serving all data within the global scientific community.

Relevant policies, organisations, activities:

- e-IRG, DG Information Society and Media, National Science Councils, OGF, FP7+, ENISA.

4.8.1.2 Normalisation Institute

If data comes from many different sources, it will need to be aligned. A normalisation institute could be set up to first contribute to standardised access across organisational and international boundaries, producing validated aggregation processes and conversion schemas – in order to achieve in the long term good overall interoperability, availability and durability of scientific data. This would be complemented by support for digital libraries and other means to take care of data curation, software curation and semantic metadata. Without these, data loses its meaning and cannot be transferred to knowledge by scientists any more.

Next steps:

- Create an enrolment mechanism for data source maintainers to use the European Grid storage facilities as a replicator to secure at least one copy of their data for free.
- Identify key data sources and fully fund their addition to the European Grid storage facilities, coordinated by a Task Force that identifies and prioritizes strategic resources.
- Fund research in replication strategies for very large database.
- Set up European repositories and digital libraries geared towards scientific software curation and serving semantic metadata.
- A normalisation institute could be set up to contribute to standardised access and aggregation.

End destination

- A complete and easily usable mirror (with affiliated metadata) of every significant data source in the world, available either real-time or with a time lag.

Relevant policies, organisations, activities:

- e-IRG, ESF, DG Information Society and Media, DG JRC, DG Eurostat, DG Internal Market, FP7+, OECD, D4Science

5 Possible Organisational Infrastructure concepts and components

It is clear that a number of the preservation infrastructure components described above are themselves archives which need to preserve digital information over the long term

and which therefore themselves require the support of that very preservation infrastructure. For example any of these components must themselves be able to be handed over to another host organisation, and the Persistent Identifiers must support such a move and resolve correctly.

An initial organisational setup could be supported by a government-level organisation, for example a component of the EU, however the commitment to provide a service for an indefinite time tends not to be popular. Therefore in the long term the responsibility could be handed over to an arms-length or consortium based organisational structure. Even this would need to be underpinned by governmental guarantee in order to provide real confidence in the infrastructure's longevity.

6 Possible Financial Infrastructure concepts and components

It seems difficult to avoid the conclusion that the initial funding to develop these infrastructure components must be provided by, for example, the EU. However given that there is also significant commercial need for digital preservation, although this tends not to be for the indefinite future, there may be options to create a self-funding set of services, especially where the service does not scale with the amount of data needing preservation. The Registry of Representation Information, the Knowledge gap manager, the Authenticity tools, the licence tool dark archive, the brokerage systems and the certification system, to name a few, do not scale with the amount of information being preserved. For example one piece of Representation Information may be used to describe 1 billion data objects. The Storage Facility on the other hand would grow with data growth, although the declining cost of storage means that this does not imply a simple relationship.

7 Possible Policy infrastructure concepts and components

There are a number of broad policies or statements of intents about preservation. Although it is not clear when or whether these will converge, it is clear that there is almost certainly be a variety of such policies for the foreseeable future. The preservation infrastructure must be able to operate in this environment.

8 Things excluded from the Roadmap

A number of preservation related activities have been excluded from this document on the basis that it is not at all clear that an infrastructure can be create to support these activities, however this must be reviewed. The list of excluded topics is as follows:

- Budgets
- Decisions of what to preserve i.e. appraisal

- Cost benefit analyses
- Access methods
- Specific domain software
- National legal aspects