Building a science data infrastructure for preservation tested in the Earth Science domain

Mirko Albani [(1)], David Giaretta [(2)] and Fulvio Marelli[(3)] , Sergio Albani[(1,4)]

*[(1)] **ESA-ESRIN***
Via G. Galilei, Frascati, Italy
Frascati, Italy
EMail: mirko.albani@esa.int
*[(2)] **Alliance for Permanent Access***
Millers Cottage, 2 High Street, Yetminster, Dorset DT9 6LF, UK
EMail: director@alliancepermanentaccess.org
[(3)] ***Advanced Computer Systems (ACS)***
Via della Bufalotta 378, 00139 Roma
EMail: fulvio.marelli@acsys.it,
*[(4)] now at **EUSC,***
Apdo de Correos 511, 28850 Torrejon de Ardoz, Madrid, Spain
EMail*: s.albani@eusc.europa.eu*

ABSTRACT

The SCIDIP-ES (**SCI**ence **D**ata **I**nfrastructure for **P**reservation – **E**arth **S**cience) project is being funded[1] by the EC (European Commission) in the framework of the FP7 program and it is about delivering long-term preservation services as part of the data infrastructure for e-Science. SCIDIP-ES aims not only to deliver generic infrastructure services for science data preservation that address the persistent storage, access and management but also to build on the experience of the ESA Earth Observation Long Term Data Preservation (LTDP) programme to favour the setting-up of a European Framework for the long term preservation of Earth Science (ES) data through the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of the generic infrastructure services in the ES domain.

The generic services will build on the already proven research prototype services from the CASPAR project. During the project these services will be evaluated and tuned in depth, using Earth Science as pathfinder, and broadly but less deeply across other disciplines linked to the Alliance for Permanent Access (APA) and ESFRI clusters. Earth Science presents an enormous challenge to the providers of data-infrastructure because it is inherently a very broad and scattered domain with completely different instruments operated by different entities, which at the moment apply different data preservation policies – or none at all. Long-term accessibility and exploitability of Earth Science data requires that also the associated knowledge needed to guarantee their understandability is properly preserved and made accessible and this requires, in addition to the application of the generic services, the availability of common policies, harmonized metadata and ontologies that will be defined in the project. This initiative is important because it will allow our society to properly preserve the digitally encoded information on which we all depend, in particular Earth Science measurements which can never be repeated and yet on which a multitude of ecological, economic and political decisions must be based in the future. The generic services will allow all kinds of data to be usable by researchers from many different domains and will enable the cost for long-term usability across disciplines to be shared supporting the creation of strong business cases for the long term support of that data.

Keywords: digital preservation, e-Infrastructure, Earth Science, science data

[1] This research is being part funded by the European Commission under contract FP7-INFRASTRUCTURES-2011-283401

INTRODUCTION

It is generally agreed that there is a data tsunami approaching. Yet the analogy is not quite right; there are many projects generating huge amounts of data each of which has its own funding, its own set of users, its own schedule, and those responsible for the data are running with all their might to cope with their own problem. Many, perhaps most, of these are large international projects, which have their own specialised infrastructure. It would be convenient if all of these could stop and wait for the generic infrastructure to be put in place, or for all of them to use the same systems, or at least for them all to agree on a common set of interfaces. Yet this is impractical.

The obvious question is then - how can a relatively small EU project position itself to be able to influence and be used by the significant mass of these projects and their users? What we wish to provide are **generic services**, but they need to be easily adaptable to fit into existing discipline infrastructures. The principle of subsidiarity must apply in order for each data producer to continue on their own courses, but commonalities must be identified and offered in such a way as to be easy to integrate into these existing specific infrastructures and applications. These services must demonstrably offer immediate benefits to repositories and science users

The **Earth Science domain** is a typical example of the above. Global changes, environmental degradation and sustainability are vital aspects that need to be understood and managed today and in future and this leads to the need to preserve a flood of Earth Science data and associated knowledge. These data, relevant services, projects' results and applications are accessible today only in a very scattered way through different providers, specialized institutes or service companies. Often, these can only be exploited by a limited community with specific knowledge on what, where, and how to search for what they need. Providers and consumers often do not speak the same scientific language and this makes the access to data more difficult in general. Petabytes of data about our planet are available today, and there are still numerous obstacles to be dealt with to guarantee their preservation and availability in the future and to facilitate their accessibility and exploitability.

SCIDIP-ES OBJECTIVES

Our aim is to overcome this problem by delivering services for long-term preservation and usability as part of the data infrastructure for e-Science. We will combine a data centric point of view - using a proven design for generic infrastructure services, for **persistent storage, access and management** - with a **user-centric point of view**, which will be built on the specific requirements coming from the Earth Science community. The former comes from leading research projects in digital preservation and the latter from the developing European Framework for the long term preservation of Earth Science (ES) data. The proven generic preservation services which were prototyped in CASPAR [1] will be made robust and scalable in SCIDIP-ES [2]. Putting the services in place is one part of the project; the other part is related to the harmonization activities in the Earth Science domain.

Long-term accessibility and exploitability of Earth Science data requires that also the associated knowledge (e.g. technical and scientific documentation, algorithms, data handling procedures, etc) needed to guarantee their understandability is properly preserved and made accessible. This can be achieved through the availability of common preservation policies, harmonized metadata and semantics/ontologies and the application of the generic services. Through the definition of common preservation policies and the harmonization in the ES domain, we will moreover boost the development of a Earth Science framework for Long Term Data Preservation facilitating interoperability among the different actors and behaving as a pathfinder initiative addressing the long term preservation of data in this challenging and sensitive domain.

The SCIDIP-ES consortium is coordinated by ESA and supported through the on-going ESA Earth Observation LTDP programme [3]. This will ensure that a critical mass of users will be created by working with ESA to aggregate the Earth Science domain. The participation to the SCIDIP-ES project of the Alliance for Permanent Access [4] will ensure the dissemination of results to the widest community of data preservation stakeholders currently existent in Europe.

APPROACH

We approach sustainability from three directions, namely (1) identifying the benefits of preserving a piece of information to make it easier to build the business case for preservation, (2) reducing the costs of preservation and (3) helping in finding the next holder in the chain of preservation, if the current holder becomes unable to provide the necessary funding or otherwise ceases to exist.

The SCIDIP-ES approach to Long Term Data Preservation is also informed by the recently published HLEG report [5], which calls for an international framework for a Collaborative Data Infrastructure. One aspect of their vision was that "*Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data and they can evaluate the degree to which the data can be trusted*".

Moreover impediments were identified in order to provide guidance for implementing this vision, and we can take address of these in the following ways. The SCIDIP-ES project will:

- Work closely with real users, in particular but not limited to the Earth Science domain, and will build on what they require. In this way, we will ensure their adoption of the infrastructure services.
- Provide an effective governance and maintenance of the services from the start and will not trying to impose a top-down system, so that we will have an infrastructure that is not too complex to work.
- Address disciplinary and cross-disciplinary strategies for metadata definition so to ensure that data can be re-used.
- Apply the subsidiarity principle – so we do not to appear to tread on researchers' toes – and take advantage of the growing need for researchers to use data from outside their own discipline, we will overcome lack of willingness of projects/funders/nations to take part and use the infrastructure services.

Interwoven, and consistent, with these points is that the business case for preservation depends on the extent to which the data will continue to be used over time. That in turn depends on three things – the trustworthiness of the data, the resolution of IPR issues and the availability of tools that enable and encourage re-use. To answer all these challenges we must address at least three stakeholders:

- For service providers, who will be for the first many years (i.e. not eternally) ourselves, the services must be extremely distributed, heterogeneous and asynchronous, which can be re-implemented over time and for which there is no single point of failure. The services are provided from outside the repositories and must be able to support the expected demand and be expandable as that demand increases to thousands of repositories and millions of users. The services must supplement capabilities of existing repositories rather than replace them.
- Data creators and data repositories must be able to create the "metadata" which the generic services need in order to work, such as data descriptions, provenance and data rights. For these we must provide toolkits i.e. applications which they use to create the metadata.
- Data users must benefit from the services with little additional effort on their part. We therefore need to integrate the services into a representative sample of the applications which they use.

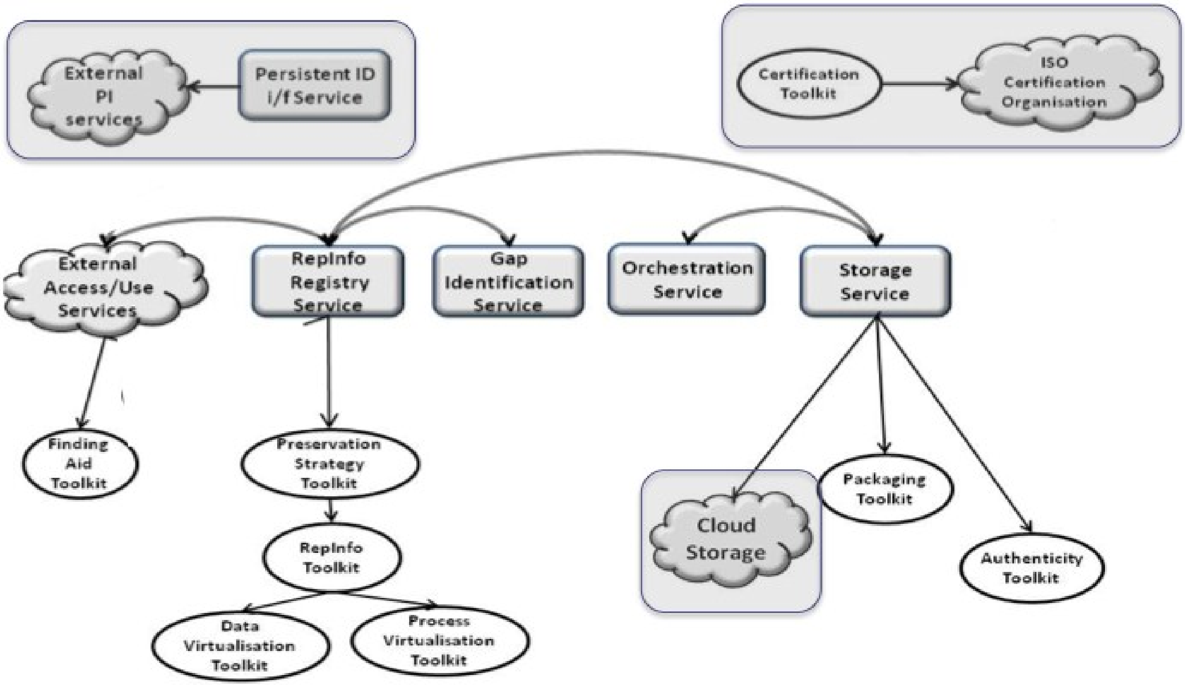
The Data Centric Approach

The CASPAR project has begun the development of techniques to create preservation research assets, which force the curators to identify, or at least to try to identify, the benefits which can arise from the use of these assets. Sustainability requires funding to be justified and the natural sequence of handing on to successive curators means that the chain of preservation is only as strong as its weakest link. It is also fundamental to identify "what" should be preserved as we have the risk, for example in the Earth Science domain, to target for preservation only a subset of the information and knowledge associated to the data that should be needed to make them understandable and exploitable in future. The services to be provided have been identified in a large body of evidence which has been collected by PARSE.Insight ([6] and [7]) from several thousand researchers, publishers and data managers across disciplines and from around the world. They spoke with almost one voice in recognising the major threats to digitally encoded information. These are summarized in the next table (left column).

Threat	Requirements for solution
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*[2]
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, for example the ability to run software in virtualized/emulated environment*
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 make it difficult to reuse data, or alternatively may not be respected in future	*Ability to deal with Digital Rights correctly in a 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*

The specific areas of the Open Archival Information System standard we will use during the project and need to focus on are those connected with the construction of Archival Information Packages (AIPs). An AIP conceptually contains all the information required for the long term usability of digitally encoded information. The assumption here is not that an organisation will 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.

[2] Representation Information (RepInfo) is a term introduced in the OAIS Reference Model [8] as the information needed by users to enable them to understand and use encoded information (i.e. data)


External PI services
Persistent ID i/f Service
Certification Toolkit
ISO Certification Organisation
External Access/Use Services
RepInfo Registry Service
Gap Identification Service
Orchestration Service
Storage Service
Finding Aid Toolkit
Preservation Strategy Toolkit
RepInfo Toolkit
Data Virtualisation Toolkit
Process Virtualisation Toolkit
Cloud Storage
Packaging Toolkit
Authenticity Toolkit

- In order to construct the AIPs, RepInfo is needed – available from the **RepInfo Registry Service** and created using ***RepInfo toolkit***. The latter is an open ended collection of tools which we divide into ***Data Virtualisation*** and ***Process Virtualisation*****.** The ***Preservation Strategy Toolkit*** helps data holders decide which of several preservation strategies to follow, based on preservation aims, costs and risk. Gaps in the RepInfo Network are identified using the **Gap Identification Service**.
- In addition we need what OAIS defines as Preservation Description Information, which largely addresses concerns about Authenticity and consists of various types of information: Reference, Context, Provenance, Fixity and Access Rights. These are dealt with in the ***Authenticity Toolkit***.
- The AIP is created by the ***Packaging Toolkit*** and stored using the **Storage Service**, which itself might delegate the bit storage to local or cloud storage services. Green issues are discussed in the context of the Storage Service since this is where we believe that the greatest gains are to be made.
- The **Orchestration Service** provides a brokerage service between existing data holders and their successors.
- The ***Certification Toolkit*** helps repositories collect evidence to submit for the ISO certification process.
- There are many identifier services which claim to be persistent. We do not propose to create another such services but rather to provide an interface, the **Persistent Identifier Interface Service**, to one or more chosen ones.

The User Centric Approach

Earth Science data, relevant services, projects' results and applications are accessible today only in a very scattered way through different providers, specialized institutes or service companies. Often, these can only be exploited by a limited community with specific knowledge on what, where, and how to search for what they need. Providers and consumers often do not speak the same scientific language and this makes the access to data more difficult in general. Petabytes of data about our planet are available today, and there are still numerous obstacles to be dealt with to guarantee their preservation and availability in the future and to facilitate their accessibility and exploitability. In particular:

- Data are not easily discoverable and accessible due to their heterogeneous nature (data coming from different sensors on different platforms like satellites, aircraft, boats, balloons, buoys or masts, or located on the land) and because they are spread all over the world and originate from different applications;
  - Assisted by the ***Finding Aid Toolkit*** and addressed by the metadata and semantics/ontology harmonization
- A common and standard approach among data providers to guarantee a harmonized and efficient preservation of Earth Science data and associated knowledge in the long term in Europe is missing; the knowledge associated to Earth Science data is moreover not defined and left to the individual provider's experience;
  - Addressed by the ***Preservation Strategy Toolkit*** and by the Common Preservation Policy
- Data are often distributed across different locations according to topic and volume, are maintained in often proprietary data formats and structures and interoperability among the different data repositories is not achieved. In particular as far as historical data is concerned there is the additional difficulty arising from the absence of systematic data curation and long term preservation procedures and tools for responding to the emerging needs to handle those data to understand climate and environmental changes over time;
- Addressed by the **RepInfo Registry** and **Gap Identification Services**, by the ***RepInfo Toolkit*** and by the Common Preservation Policy and interoperability analysis and consolidation
- Different data access policies (which may as well change over time) are associated to different datasets precluding wide access and use;
  - Addressed by the ***Packaging Toolkit*** components which allows users to capture the Access Rights associated with the data, and associate those rights permanent with the data.
- Even when accessible, data cannot easily be integrated or inter-operated, especially in a multi-discipline environment for more complex data usage, and used because of missing tools, specific algorithms, etc. needed to interpret the data correctly.
  - Addressed by **the RepInfo Registry Service** and the ***RepInfo Toolkit***.
- Sufficient computing resources, tools and algorithms to obtain the results in a reasonable time are lacking and available applications and resources are often not able to handle very large and spatially distributed data sets.
  - Not addressed by our services but rather by the underlying e-Infrastructure services which are being put in place.

As we see SCIDIP-ES will not be able to address all the issues listed above but, building on the existing OAIS model and following the INSPIRE directive [9] amongst others, will address enough of them to bring a coordinated approach addressing European Earth Science data and associated knowledge preservation. Very importantly this will facilitate their exploitability in Earth environment research and monitoring activities in the future.

The generic services we offer are those described above. Moreover good prototypes are available for many of these services from the recently completed CASPAR project which has been proven to be usable across disciplines and to have proven effectiveness in. Additional services will be developed starting from the ones coming from other projects like SHAMAN [10]. In addition there are ideas from domains which are already very advanced in terms of data management and access such as the Earth Observation domain, which is part of the broader Earth Science domain; such services will be consolidated in the project and generalised in order to allow utilization in different application domains.

Bringing together the bottom-up and top-down approaches, the proven generic services developed in SCIDIP-ES will be tailored to the Earth Science domain specific needs. This will allow Earth Science to enter in a new era where environmental research and monitoring can be carried out by better and more easily accessing and exploiting the huge amount of historic and modern information available in Europe. To this end, interoperability aspects among ES data providers and repositories in Europe will be analysed and the architecture, the business case and the governance model of an European Earth Science LTDP infrastructure will be defined in the initiative. The harmonization of rights and Intellectual property frameworks for the access to Earth Science data and associated knowledge (i.e. data access policies) will also be analysed and addressed in line with EC directives such as INSPIRE and the "GEO-GEOSS Data

Sharing Principle [11]. The goal is to achieve sustainability in the long term, according to the sustainability models adopted for example in the ESFRI projects and consolidated in the initiative and to facilitate access to data for users, while respecting data providers' policies where necessary.

EARTH SCIENCE ACTIVITIES

It is important to point out that the "knowledge" associated to Earth Science data is not well defined at the moment. It is not clear in fact what kind of additional information need to be preserved in the long term in addition to the primary data to allow the future exploitability of the data themselves. The picture is made more complicated by the fact that it is also not known at the moment how data may be used by future scientists and researchers and therefore the preservation of information that seems not useful today might be of great importance for future generations of users. Consider for example the preservation of satellite optical images acquired over cloudy areas. These kinds of images were not considered useful as the primary target of the instrument acquisition, the Earth Surface, was covered by clouds and therefore not visible in the collected images. Space agencies were thinking of removing the cloudy optical data from their archives, also to reduce costs, but mentioning this intention at a symposium with scientific users they were accused by the audience of being "criminals" as they were risking to delete fundamental information needed to analyse and study the clouds themselves!! In addition to the common preservation policies therefore, we will also define, to the best possible knowledge at this time, what information and knowledge associated to each category of Earth Science data need to be preserved to guarantee that the data can be used and exploited in the future according to the current utilization of the data themselves but also trying to guess what other information, today not considered relevant, could be considered worth to be maintained properly. The definition of common preservation policies in the Earth Science domain defining "how" the data should be preserved in the long term and made accessible together with the definition of "what" precisely should be preserved for each different instrument category marks a fundamental step forward as today all of this is left to each organization in isolation

Last but not least, the harmonization of metadata, semantics and ontologies and their utilization in the services developed within the initiative, opportunely integrated within ES providers infrastructure, will facilitate the discovery and access by users to ES data that are heterogeneous nature and spread all over the world. This initiative, acting as a pathfinder in the sensitive and multidisciplinary Earth Science domain, will generate models and results which we believe are adoptable and exploitable in other scientific and non-scientific domains.

The basic principles of a common European distributed archiving concept aiming at the creation of an interoperable network of archive centres possibly reusing infrastructure of the different entities (as a single archive for example consisting of Federated archives (OAIS), i.e. separate infrastructures with common access mechanisms) in the long term will be defined in the initiative and the governance and financial aspects and governance models of such European Earth Science LTDP Framework and Infrastructure will be analysed and addressed. Cooperation programs based on standardized and certified services (share of data archives, archive transfer on demand or in case of a purge alert, coordination of re-processing schemes, format adoption/conversion, etc.), brokering of information and agreements about available resources (e.g. obsolete hardware) and about sharing responsibility for datasets (e.g. agreeing to maintain additional copies or agreeing to take over the custody of datasets) will be part of the basic principles.

Interoperability aspects among the different categories of Earth Science data and related providers will also be analysed. Needs and gaps will be identified with the purpose to pursue harmonization to the maximum extent within and among the different Earth Science data domains to minimize costs and maximise interoperability and synergies. The tools/services/components to be enhanced (or developed) and the upgrades needed in the different infrastructures (e.g. archives and access systems) to ensure interoperability within single Earth Science data domains and across them will be identified.

Impact analysis on the current infrastructure of the different initiative participants in the different data domains will be also performed in light of the Earth Science Infrastructure principles. The architecture of this European Infrastructure, based on the upgrade and federation of existing components and on the integration of the generic services developed in the project will be defined.

In addition to technical infrastructure and capabilities, the long-term management of Earth Science data requires organizational sustainability to provide continuing stewardship to address the risks to scientific data and support their use by future communities. Providing sustainable infrastructure for the preservation of scientific data requires organizational commitments, capacity, structures and plans for data stewardship that are consistent with the missions of the organizations that accept the responsibility to serve in data stewardship roles. Alternative approaches to attaining organizational sustainability for interdisciplinary human dimensions and polar data are discussed in terms of recent recommendations for organizational sustainability to foster digital preservation. To this end SCIDIP-ES will also define the governance and organization model of the ES infrastructure with the goal to achieve sustainability in the long term, according to the sustainability models adopted for example in the ESFRI projects, and to pursue a maximisation of the open access to data for users respecting individual provider's data policies where necessary.

CONCLUSION

The SCIDIP-ES initiative is based on a solid body of research into proven generic techniques for digital preservation. Critical measures of success are

1. the continuation of the services beyond the end of EU funding, as part of the broader European (and perhaps global) e-Infrastructure
2. have a "critical mass" of users – initially in the Earth Science community

During and after the EU funding phase the use of the services are expected to grow within and beyond the Earth Science community.

The benefits will include

1. sharing the costs of preservation
2. improving the ability of individual repositories to preserve their holdings by supplementing their existing systems, making it easier for them to achieve certification [12]
3. improve the usability of data holdings across domains and thereby broadening the user community, as foreseen by the High Level Expert Group.

REFERENCES

[1] CASPAR project – see http://www.casparpreserves.eu .(The project was supported in part by the European Commission under contract IST-2006-033572)

[2] See http://www.scidip-es.eu

[3] Long Term Preservation of Earth Observation Space Data: European LTDP Common Guidelines, Issue 1.1, Ground Segment Coordination Body, 30 September 2010.

[4] See http://www.alliancepermanentaccess.org

[5] Report of the High Level Expert Group on Scientific Data available from http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707, with the associated press release at http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

[6] PARSE.Insight web site at http://www.parse-insight.eu and in particular the Roadmap http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf .(The project was supported in part by the European Commission under contract FP7-2007-223758)

[7] PARSE.Insight general survey report http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[8] http://public.ccsds.org/publications/archive/650x0b1.pdf or later version. At the time of writing the revised version is available at http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf or elsewhere on the CCSDS web site http://www.ccsds.org

[9] INSPIRE Directive – see http://inspire.jrc.ec.europa.eu/

[10] SHAMAN project - website at http://shaman-ip.eu/shaman/

[11] GEO-GEOSS Data Sharing Principles see http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf and the Implementation plans at http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf

[12] Audit and certification of repositories – see http://wiki.digitalrepositoryauditandcertification.org and http://www.alliancepermanentaccess.org