eInfrastructures and Digital Libraries… the Future

Wim Jansen

DG Connect, EC

Prof. Roberto Barbera

Dipartimento di Fisica e Astronomia &
INFN Sezione di Catania, Italy

Prof. Yannis Ioannidis

University Of Athens
Greece

Michel Drescher

European Grid Infrastructure (EGI),
The Netherlands

Dr. Norbert Meyer

Poznań Supercomputing and Networking
Center, Poland

Antonella Fresa

PROMOTER, Italy

Nick Poole

Collections Trust, UK

Prof. Matthias Hemmje

Fernuniversitat Hagen, Germany

Prof. Peter Stanchev

Kettering University, USA

1 Setting the scene

1.1 Background (Yannis Ioannides)

According to the DELOS Reference Model on Digital Libraries, a Digital Library is a (potentially virtual) **organization** that comprehensively collects, manages, and preserves for the long term rich **digital content** and offers to its **user** communities specialized **functionality** on that content, of measurable **quality**, and according to prescribed **policies.** Furthermore, a Digital Library System is a software system that is based on a (potentially distributed) **architecture** and provides all functionality that is required by a particular Digital Library.

For the most part, Digital Libraries and even Digital Library Systems are still constructed in an ad hoc fashion, in the context of distinct and often isolated efforts that start from scratch to build systems supporting the particular needs of the Digital Libraries concerned. Nevertheless, there is substantial commonality among many of them that should be realized within the data infrastructure layer of the eInfrastructure ecosystem, leaving only specialized services to individual implementation efforts on top. Although data infrastructures are now conceptually well established, their implementation maturity is lagging behind. This is due to one of the foundational

adfa, p. 1, 2011.

parameters, however, that appears only in data infrastructures (and Digital Libraries): digital content, be it digitized versions of books and articles, or born-digital content, such as structured, semi-structured, and unstructured data, ontologies, and multimedia. And that's exactly what makes data infrastructures so much more challenging than the rest. Digital content can be conceived at several levels of syntactic and semantic abstraction, making it rather complex to extract, store, process, analyse, and interoperate automatically. There is currently a substantial effort to overcome these difficulties and develop technologies that will make data infrastructures a reality. To be really successful and change users' everyday life, data infrastructures should provide really sophisticated services that address complex user needs. They should not confine themselves to the bare minimum of file input/output simply augmented with some auxiliary functionality. They should take advantage of the advanced data management technology that has been developed by the database research community in the last 40 years and organize digital content so that rich data services can be provided on top of it. That's where the future of Digital Libraries lies!

1.2 Milestones in development (Peter Stanchev)

There are several initiatives to connect research article with the data. The term Enhanced Publication means a new type of publication whereby researchers link directly through to supplementary data.

In July 2012, the European Commission issued a recommendation on access to and preservation of scientific information. e-InfraNet and the Knowledge Exchange initiative organized a workshop bringing together e-Infrastructure providers and information experts from the Europe member states to explore which actions were required to fulfill this recommendation.

The OpenAIRE project supports the implementation of Europe's open access (OA) policies as outlined in the European Research Council's Guidelines for Open Access and the European Commission's Seventh Framework Programme (FP7) Open Access Pilot. In its efforts to reach out to research institutions, researchers, project coordinators and publishers in the individual European countries, OpenAIRE is facilitated by a network of National Open Access Desks (NOADs). The NOADs also provide support to institutions in developing their OA policies: in implementing the European Commission's OA Pilot and the ERC's Guidelines on OA, in building synergies within institutional OA policies, and in making repositories and OA journals compliant with OpenAIRE's requirements for metadata harvesting.

OpenAIREplus project aims at bridging the missing link between the research articles, the data and the project funding. Building on the OpenAIRE portal and OpenAIRE compatible repositories, OpenAIREplus harvest multiple heterogeneous sources and by using machine power find the links between articles, data and projects.

The EUDAT project brought together data infrastructure providers and practitioners to discuss current data infrastructure challenges and solutions, with a focus on interoperability, cross-disciplinary and cross-border collaboration.

The Research Data Alliance was created. The purpose is to accelerate international data-driven innovation and discovery by facilitating research data sharing and

exchange, use and re-use, standards harmonization, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards, and other deliverables.

2 Examples from the cultural heritage domain

2.1 EUDAT: A New Pan-European Cross-Disciplinary Data Infrastructure for Science (Norbert Meyer)

In recent years significant investments have been made by the European Commission and European member states to create a pan-European e-Infrastructure supporting multiple research communities. As a result, a European e-Infrastructure ecosystem is currently taking shape, with communication networks, distributed grids and HPC facilities providing European researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. However, the accelerated proliferation of data – newly available from powerful new scientific instruments, simulations and the digitization of existing resources – has created a new impetus for increasing efforts and investments in order to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation. EUDAT is a pan-European initiative that started in October 2011 and which aims to help overcome these challenges by laying out the foundations of a Collaborative Data Infrastructure (CDI) in which centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities.

During the first two years of the project, EUDAT has been reviewing the approaches and requirements of a first subset of communities from linguistics (CLARIN), solid earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), and shortlisted four generic services to be deployed as shared services on the EUDAT infrastructure. These services are data replication from site to site, data staging to compute facilities, metadata, and easy storage. Additional services are currently being considered for discussion and inclusion in the second development round. They are concerned with issues related to data annotation, handling of (near) real-time data, crowd-sourcing, and web-services.

The services being designed in EUDAT will thus be of interest to a broad range of communities that lack their own robust data infrastructures, or that are simply looking for additional storage and/or computing capacities to better access, use, re-use, and preserve their data. Although EUDAT initially focused on a subset of research communities, discussions with other research communities – belonging to the fields of environmental sciences, biomedical science, physics, social sciences and humanities – have already begun and EUDAT aims to associate these communities to the design of the infrastructure and its services.

2.2 DCH-RP[1] project (Antonella Fresa)[2]

The result of the digitisation initiatives carried out by memory institutions (museums, libraries, archives) in Europe and world-wide have produced a large amount of cultural content, which is continuously growing.

The existing services – e.g. the metadata aggregators, such as Europeana and the national cultural portals -, even if important to mobilise attention and resources on the general theme of the digitisation of cultural heritage, have limitations that do not allow to unlock the whole of this potential. On the other hand, the research infrastructures (e.g. DARIAH) are currently missing most of the cultural heritage data, including the data that are hold by local institutions.

The idea of a digital cultural heritage e-infrastructure is to set-up a "common pot", where institutions can deliver safely their content, which can be seen as a "continuum" by the users. Further, the content that are ready to be open can then be used to develop and demonstrate use cases to encourage more institutions (both national and local one, also small institutions) to adhere to the initiative.

The digital cultural heritage content is made of several different kind of information: data (2D images and 3D models), metadata, publications, digital exhibitions, virtual reconstructions.

Even if growing very rapidly, the actual size of this content, if measured in byte, is still very much smaller than the amount of data produced by the experiments and the observations of the "hard sciences".

However, its value is very much high, also because the content related to the digital cultural heritage is produced with a very intensive human work, which makes it expensive and unique. Therefore, preservation has the highest priority for the digital cultural heritage domain and the e-infrastructure services envisaged for cultural heritage and humanities should include preservation features, ranging from long-term to short-term storage. In addition to the storage resources, also the computing resources of the research e-infrastructures (both grid and cloud) can serve the preservation needs. An interesting experiment is the use of the grid to perform regular check-sum test, to monitor any damage or corruption to the data.

Further, the digital cultural heritage content is composed by information that is strongly linked. For example, the individual cultural object needs to be studies with respect to its context (that is made of information possibly coming from different data bases) and in the frame of the collections it belongs to. The improvement of the search technologies, in particular the application of semantic technologies, is very promising in this field. More generally, the linked (open) data movement is very much discussed in the cultural sector as a frontier to be reached to empower the digital cultural content with more links. Features to manage semantic search and linked open data are necessary components of the digital cultural heritage e-infrastructure.

---

[1] Digital Cultural Heritage Roadmap for Preservation, www.dch-rp.eu

[2] Contacts: Antonella Fresa, Technical Coordinator, fresa@promoter.it; Borje Justrell, WP3 Preservation Roadmap Leader, borje.justrell@riksarkivet.se; Rosette Vandenbroucke, WP4 Case Studies and Best Practice Leader, rosette.vandenbroucke@gmail.com

Even if not strictly a technological matter, the issue of openness of digital cultural content is another main challenge that the sector is approaching when moving online. To this regard, it should also be noted that the digital cultural heritage content is of great interest for many different targets: the research (both humanities and other sciences, e.g. archival data are currently offered for investigation to medical researchers and economists), educational products and the creative industries. Often, the partnerships for the exploitation of cultural data (both use and re-use of data) see public and private organisations working together, with the need to combine commercial and not-for-profit approaches. For this reason, digital cultural content needs to be accessed differently by each target, and therefore, authentication and authorisation mechanisms are important to be put in place efficiently by the e-infrastructure.

3 Examples of data infrastructures

3.1 European Grid Infrastructure (EGI)[3] (Michel Drescher)

EGI.eu is a not-for-profit foundation established under Dutch law to coordinate and manage the European Grid Infrastructure (EGI) federation on behalf of its participants: National Grid Initiatives (NGIs) and European International Research Organisations (EIROs). National Grid Initiatives are in turn organisations that coordinate and manage the national resources comprising distributed computing and data infrastructures through its members, the Resource Centres.

The EGI is continuing the work and delivery of services that have been provided for the European and worldwide research communities that traditionally were predominant in the field of academic distributed e-Infrastructures: Researchers from the disciplines of High Energy Physics, Astronomy, Astrophysics and Astro-particle Physics, Life Sciences, Earth Sciences, Computational Chemistry and other multidisciplinary sciences.

Therefore, the traditional e-Infrastructure delivered to these "hard sciences" (as coined by other research communities) was geared in their capabilities to match the needs of exactly these disciplines – that is, batch-orientated submission of descriptions of relatively similar computing instructions (often referred to as "compute jobs").

Being founded in 2010, EGI started to realign itself into driving its evolution by requirements coming from the user communities it is supporting, so as to explicitly provide a common foundation upon which the digital European Research Area can be brought online: the EGI ecosystem. Supported initially through the EGI-InSPIRE project, EGI focuses on delivering a production-quality e-Infrastructure to the supported research communities. This is not to diminish the importance of the other strategic activities: without these complementary pillars, EGI would not be able to deliver an operational infrastructure at its best potential. This strategy (http://go.egi.eu/EGI2020) describes three pillars of the EGI ecosystem:

[3]Contacts: Michel Drescher, Technical Manager, Michel.Drescher@egi.eu

Pillar 1 – Operational Infrastructure: The Operational Infrastructures provides the technical ICT foundation of the EGI e-Infrastructure by providing a distributed, federated service platform for access by end users. Depending on the needs of the targeted research community, EGI offers federation and operational services, Cloud Infrastructure services, or collaboration services.

Pillar 2 – Community and Coordination: Integral to delivering a pan-European e-Infrastructure are services around social aspects of a large and complex e-Infrastructure. What is often described as "connecting people" includes community building, development of human capital, coordination, communication, and last but not least strategy and policy related activities across the entire EGI community.

Pillar 3 – Virtual Research Environments: Virtual Research Environments (VRE) are defined as the complete and inclusive work environment that is owned, deployed, managed and used by one or more closely related research communities. This definition includes ICT resources that are entirely remote and external to EGI as well as EGI resources that are, or will be, integrated into potential VREs. Support for Virtual Research Environments includes infrastructure services such as deployment and hosting of Community Platforms on top of EGI resources, but also consultancy and technical services for existing and new community services.

EGI's operational infrastructure is organised into three distinct platforms that are for the most part owned and operated by EGI or, in case of some services that are part of the Collaboration Platform, by selected and trusted external partners. These three platforms – EGI Core Infrastructure Platform, EGI Cloud Infrastructure Platform and EGI Collaboration Platform - provide a multi-level generic infrastructure that is not in any way in predisposition towards any research community. However, the combination allows deploying any number of Community Platforms on top of it. Community Platforms are defined as a specific set of services in a suitable configuration, deployed remotely as part of the EGI Operational Infrastructure, serving a specific community's ICT needs.

Currently spanning over 340 Resource Centres in over 50 countries in Europe and worldwide[4] providing over 360k CPU cores and 235 PB Disk and 170 PB Tape storage) EGI is in a position to support an e-Infrastructure that is tailored to the needs of the Digital Cultural community on a European scale. Leveraging the self-service infrastructure deployment model introduced with the EGI Cloud Infrastructure Platform this model allows integrating existing digital data repositories into a virtual digital library e-Infrastructure that overlays the resources provided through the EGI production infrastructure: Digital artefacts may be preserved in a mixed mode (both local and in the EGI cloud), while computational services may be completely deployed on EGI's Cloud infrastructure.

[4] These numbers include integrated resource providers that are indirectly part of EGI

3.2 CHAIN-REDS project[5] (Roberto Barbera)

E-Infrastructures are being built since several years both in Europe and the rest of the world to support diverse multi/inter-disciplinary Virtual Research Communities (VRCs) and a shared vision for 2020 is that e-Infrastructures will allow scientists across the world to do better (and faster) research, independently of where they are deployed and of the paradigm(s) adopted to build them.

E-Infrastructure components can be key platforms to support the Scientific Method, the "knowledge path" followed every day by scientists since Galileo Galilei, in many aspects. Distributed Computing and Storage Infrastructures (local HPC/HTC resources, Grids, Clouds, long term data preservation services) are ideal both for the creation of new datasets and the analysis of existing ones while Data Infrastructures (including Open Access Document Repositories – OADRs – and Data Repositories – DRs) are essential also to evaluate existing data and annotate them with results of the analysis of new data produced by experiments and/or simulations. Last but not least, Semantic Web based enrichment of data is key to correlate document and data, allowing scientists to discover new knowledge in an easy way.

So far, however, it has been difficult for researchers (not mentioning the "citizen scientists") to correlate papers to datasets used to produce them and to discover data and documents in an easy way. In this contribution we present the "Coordination & Harmonisation of Advanced e-Infrastructures for Research and Education Data Sharing" (CHAIN-REDS) project and its activities to foster and support worldwide data infrastructures.

CHAIN-REDS has devised a program for data infrastructures that aims to:

- Identify standards to easily gather and access both OADRs and DRs;
- Build a demonstrator to easy visualise and access OADRs and DRs;
- Correlate OADRs and DRs to create linked data and discover new knowledge through semantic enrichment of metadata
- Promote Data Infrastructure standards and identify new OADRs and DRs from regions addressed by the project (Africa, Middle-East and Gulf Region, Latin America, China, India, Far-East Asia);
- Populate the demonstrator with these new repositories, add them to the semantic enrichment tool, and set-up at least two use-cases from different domains

The various elements of the program will be discussed with particular focus on the Knowledge Base and its Search Engine on Linked Data.

[5] Contacts: Federico Ruggieri, Project Director, federico.ruggieri@roma3.infn.it; Roberto Barbera, Technical Coordinator, roberto.barbera@ct.infn.it

4 Mixed scenarios

4.1 Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN)[6] (Matthias Hemmje)

The SHAMAN Integrated Project has delivered a long-term next generation digital preservation (DP) framework and corresponding application solution environments for analysing, ingesting, managing, accessing and reusing information objects and data across digital libraries and persistent archives. Three prototype application solutions were built on the basis of this framework environment, testing and validating outcomes in the domains of scientific publishing and parliamentary archives; industrial design and engineering; and finally for scientific applications within eInfrastrcutures. Within the SHAMAN DP infrastructure, the core functions are organized along the SHAMAN reference architecture. Using this architecture the project has created a software framework and application development environment supporting the creation of test-beds of Digital Preservation support infrastructures and services. The core services of the SHAMAN framework are constructed by integrating Data Grid (DG), Digital Library (DL), Persistent Archive (PA), Context Representation, Annotation, and Preservation (CRP) as well as Deep Linguistic Analysis (DLA) and corresponding Semantic Representation and Annotation (SRA) technologies for simple and connected data types establishing, document, media, CAD, and scientific data-, knowledge-, and information collections. This has provided functionality and conceptual foundations for the long-term unification of knowledge preservation and analysis across eInfrastrutures for domains within a distributed grid-based infrastructures. The test-bed environment has been demonstrated and validated in a user and impact-oriented way by three complementary real-life use scenarios drawn from the libraries and parliamentary archives domain, the engineering domain, and the scientific domain.

The SHAMAN DP infrastructure meets the growing need in the European Union for the management of technology evolution for preservation environments in dealing with distributed data sources, distributed data curators, or distributed users. Therefore, SHAMAN's outputs are providing the ability for users to assemble a shared collection whose records reside in multiple types of storage systems, at multiple institutions, located in multiple nations, regions or locations. It supports the long-term preservation of digital entities through mechanisms which will manage the authenticity of massive data collections that are written to archival storage systems.

The above capabilities are urgently required throughout the European Union as well as on a global scale for collaborative research and development including the publication of data in a variety of media, support for design, engineering, and manufacturing as well scientific data, e.g. the federation of real-time data from sensor systems, and the management of technology evolution required for preservation environments. The ability to manage such data over the long-term is at the heart of the

[6] http://shaman-ip.eu/

European economic and research agenda as well as required for legal and statutory requirements for European Parliamentary and other archives.

The SHAMAN project differs from related European-funded projects in the way that it defines the entire preservation information context sufficiently well such that the records can be migrated into an independent preservation environment without loss of authenticity or integrity. This requires migrating not only the records, but also the characterisation of the preservation context itself.

To achieve its objectives, the project has integrated developments in the areas of digital libraries, data grids, and persistent archives to implement a demonstrably complete preservation environment. This includes a demonstration of automating archival processes using data grids; the development of constraint-based collection management systems which extend the state-of-the-art in data grid technologies; and the extension of this on a prototype basis to support scientific and engineering data.

The development of a stable and reliable preservation environment is strongly driven by the desire to support infrastructure independence, the ability to preserve digital entities as a collection, and the ability to migrate the collection to new choices of storage and database technologies. SHAMAN has achieved this by implementing the basic control mechanisms required to manage distributed environments, as follows:

- **The use of data grid technologies offers a future proof Digital Preservation strategy for multiple organisations and communities.** This strategy is build on the use of distributed data grid technologies to manage and administer replicated copies of digital objects over time. A primary component is the use of data grid middleware as the core data management technology which meets the currecnt requirements in supporting preservation requirements. While data grids have been successfully used for a number of preservation projects in the European Union and the United States, there are a number of additional achievements which have been reached by SHAMAN in order to enable administrators to successfully manage and administer secure digital preservation environments for production. These include the ability to extract digital entities from their creation environment, the provision of technologies which can be used to manage authenticity and integrity over time, and the provision of technologies which will enable users to discover, render, and reuse the data.

- **The use of basic control mechanisms in a distributed environment to introduce transaction capabilities to Digital Preservation services.** Typically digital preservation services have focused on OAIS defined metadata attributes, without addressing the broader requirements for transaction services, which may be used to visualise or analyse the content. SHAMAN has implemented a resource virtualisation, or the standard services to enable jobs to be executed across independently managed computational and storage resources. In addition to providing the basic functionalities of moving data around and enabling the use of high-level

data analysis pipelines, the result has also increased the capability of the digital preservation community to work together.

- **The use of digital library services to provide access to data held in preservation environments.** Typically, preservation environments have not addressed the critical aspects of discovery, analysis and reuse of data. Once data is registered in a preservation environment, it is important to enable users to discover relevant data. Thus, the project provides the basic infrastructure to meet both the needs of administrators seeking to archive data, and those of ordinary users who need to access the data after archiving. For the administrator, the project provides the tools to ensure the authenticity of data, to make certain it can always be identified relative to the context in which it was created. This is done by means of producer/archive ingest workflows, application of the Metadata Encoding Transmission Standard (METS) to label binary objects, and the use of tools to extract data from a variety of sources. This will enable users to discover data across shared collections, even if it is kept in data grid containers and otherwise inaccessible.

- **Provision of a model for capturing, representation and management of context**. The implementation of preservation processes requires knowledge about the organisational policies, along with legal/societal requirements, along with known properties of digital objects and the effects of operations carried out to them. Accordingly, SHAMAN has employed a sophisticated model for the representation and management of context. This context knowledge is exploited for the management of production, working and preservation processes, using reasoning mechanisms to automate their enactment, coordinating their execution in a flexible, robust way. Given the low state of automation in digital preservation and the lack of machine-interpretable semantics, the SHAMAN context model provides an advanced basis for the future management of digital objects.

In summary this means that SHAMAN has provided an infrastructure, reference architecture and framework that enables the cross-fertilisation between the Data Grid, Digital Library, and Digital Preservation domains that can now be re-used to foster further synergies and advances in all three areas. The integration of digital library, data grid, and digital preservation technologies as set out in SHAMAN's exemplary application domains provides a method and reference implementation that can now be re-used and extended to unleash further cross-fertilisation possibilities, as each eInfrastructure community has its own unique, but similar, requirements that can build on SHAMAN results.

4.2 INSIGHT into issues of Permanent Access to the Records of Science in Europe (PARSE.Insight)[7] (Matthias Hemmje)

The growing multitude of digital resources forms the basis of the intellectual capital of European research. Mining of further information from these resources and allowing new generations of researchers to "stand on the shoulders of giants" is the very essence of research. These digital resources must persist and remain findable, accessible, and understandable. Data re-use (by users in a different discipline, for example) may happen immediately the data is produced or may not happen for an extended period of time. The same techniques for preservation of data assets support contemporaneous (re-)users as well as the interests of future generations.

There is a very real risk that much of the scientific data and documentation that exists may be lost to future generations unless permanent access is secured. PARSE.Insight was focused on the infrastructure needed to support persistence and understandability of these key assets over the long term. As noted in the Open Archival Information Systems (OAIS) Reference Model (ISO 14721), when one talks about long term preservation, long term "is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community", which could be just a few years.

The advent of e-Science has deeply modified the research process. The century-old cycle of reading and writing scientific publications as the only medium of scientific exchange has evolved into a multitude of digital resources which form the intellectual capital of European research. These new opportunities are fostering multi-disciplinarity and accelerating the life-cycle of research, enabling the fast re-use of information crucial to scientific investigation. At the same time, while we can [and some branches of mathematics still do] still read articles of centuries ago, most scientific disciplines are effectively risking the entire capital of European research: no coherent or concrete efforts are being made to preserve the digital records of European science.

There is a real risk that our scientific records will not be findable, accessible and understandable over the medium and long term, or -in some cases- even the short-term. European science therefore risks of impairing its competitiveness as there might be no proverbial (digital) "shoulders of giants" to stand on.

[7] http://parse-insight.eu/

PARSE.Insight has highlighted this situation and concentrated on roadmapping the parts of the e-Science infrastructure needed to support persistence and understandability of the assets of EU research.

PARSE.Insight has produced such a roadmap, bringing together national, European and global thinking. A detailed inventory of existing and planned preservation support has been gathered. Comparing the roadmap to the inventory has identified the gaps in the research arena that the Infrastructures programme can help address. However, permanent access raises new challenges which need new ways of analysing potential impact; in addition to the roadmap itself, PARSE.Insight has provided and a tool to support future iterations of such an impact analysis as well as an interactive map of its actors and stakeholders.

The PARSE.Insight project had been performed by members of the European Alliance for Permanent Access (APA, http://www.alliancepermanentaccess.eu/). APA is a unique cross-sectoral coalition of major stakeholders in science and scientific information, created in response to the work of a high-level task force initiated during the Dutch Presidency of the EU in 2004.

In addition to the PARSE.Insight consortium members, APA members currently include European Science Foundation, Centre National d'Etudes Spatiales, Centre Informatique National de l'Enseignement Supérieur, the UK's Joint Information Systems Committee, the British Library and the National Archives of Sweden.

At a practical level, APA members have already been and are still active in a number of relevant key projects for long-term preservation and access such as CASPAR, DRIVER, PLANETS, SHAMAN and DPE as well as the still ongoing APARSEN and SCIDIP-ES. The involvement of commercial partners in some of those projects, particular those in the storage and ICT industries, such as IBM, Microsoft and SUN, further expands the sphere of influence of APA and its members.

APA Members are also participating in development of a new generation of scientific digital repositories which takes fully in account the European effort for building e-infrastructures such as GENESI-DR and EuroVO-AIDA, covering Earth Science and Astrophysics communities (both projects, submitted to the first FP7 Research Infrastructure call, are at this moment in negotiation).

However, projects by definition have a limited focus and time frame. Therefore, APA provides a conceptual and strategic framework that organises and consolidates these individual efforts and fills possible gaps between them. This ensures a unified, Europe-wide approach to the issue whilst working closely with relevant projects undertaken outside Europe. APA therefore has adopted a strategic work plan to help consolidate its role at national, European and international levels. The work plan is a framework for actions that will help generate critical mass and accelerate the progress that APA can make over the mid-term which will help to multiply the investments made by the EU many-fold by helping to align the investments within individual nations.

Within this context, the achievements of PARSE.Insight can be expressed and motivated in terms of six key results. APA is committed to engagement with a wide range of scientific communities, and so the first achievement follows as a foundation on which the rest of the project's work rests:

> ***Insight*** *and understanding into the capabilities and practices within the various research communities, leading to the ability to share and capitalise on best practices as well as understanding the impact that e-Science is having on the research communities that it is serving and on the long-term availability of research data.*

The capabilities and practices cover current behaviour and priorities, common ground and differences, understanding of what is possible and what is not, and why. This refers to the current situation within the research communities; of course it is also necessary to take account of planned activities to advance the state of digital preservation, hence the need for the second achievement:

> *Bringing together for the first time an* ***inventory*** *of current and planned research and development relating to e-infrastructures and permanent access, regardless of the funding mechanism, at a national, European or global level, leading to insight into areas where intervention may be required.*

This has been constructed from information gathered from a very wide range of stakeholders including funders, producers, curators/preservers and users from a wide range of disciplines. In this way APA provides access to a very wide range of disciplines, some of which are represented within the consortium but most of which are outside.

These broad surveys of PARSE.Insight have been supplemented by particular case studies. These are in-depth studies which had required significant and extended interactions, and came from within APA where the consortium had direct access.

From these, a process of analysis and synthesis has lead to the third achievement:

> *A **Roadmap** for a support e-infrastructure for maintaining long-term accessibility and usability of scientific and other digital information in Europe.*

An initial draft of the roadmap has been derived from a synthesis of the many roadmaps which already existed. This draft has provided a structure for the way in which the inventory was captured.

The roadmap was produced with dependencies, showing the milestones on the way to achieving the unified infrastructure for preservation and permanent access. The roadmap include topics such as storage technologies, software systems, support components such as registries, persistent identifier systems, types of data and documents, repository support, access and interoperability requirements.

Some of the milestones of the roadmap relate to work already planned or under way in the meantime, and covered in the inventory, while others relate to work that is still needed. W.r.t. the work still neede, the fourth achievement is:

> *Identification of **gaps** in the existing and planned infrastructure.*

The **gap analysis** is the central achievement from PARSE.Insight, generated by comparing the roadmap with the inventory of existing and planned capabilities. It allows the EU, and others, to focus resources where they are most needed in order to develop the full preservation e-infrastructure required.

These gaps relate to specific sectors or communities, or are common across sectors. An example of the latter is, for example, a common digital object identifier with a guaranteed sustainable future. In any case, given that the roadmap represents an ideal, it supports it with evidence in the form of impact analyses of the developments entailed by the roadmap. These analyses allow to debate about the costs and benefits of undertaking the developments, or of not doing so, and are useful to policy makers and programme managers at national and international levels. Thus another outcome follows:

*A framework for **impact** analysis, based on these insights, leading to better informed investment decisions and sustainable e-repositories for scientific records.*

One particular need, common across all sectors, that can be pinpointed now and on which the PARSE.Insight has achieved results, is for a common and accepted method for audit and certification of trusted digital repositories. Such a system allows, e.g, funders to judge whether they are funding the right repository service providers. It also allows best practice to be identified in different domains, thereby allowing those best practices to spread.

The PARSE.Insight project has fed into international research and development of audit and certification methodology and implementations as well as the insight gained into different domains, so providing a valuable overview and check on the process leading to the method for audit and certification in these domains. The corresponding outcome is:

*Progress in the development of an international process for **evaluating** the sustainability and trustworthiness of digital repositories, and identifying best practice.*

All above activities should be considered to focus to the common, higher level objective, i.e. to lead towards a sustainable cross-sectoral approach to preservation services for eInfrastructures.

4.3 SCIence Data Infrastructure for Preservation – Earth Science (SCIDIP-ES)[8] (Matthias Hemmje)

The aim of the currently ongoing SCIDIP-ES initiative is to address Data infrastructures for e-Science, delivering services for long-term preservation and usability as part of the data infrastructure for e-Science. However the effectiveness of preservation services is difficult to prove – except by waiting a long time and even if they are effective, users will not automatically use them. SCIDIP-ES combines a top-down, data centric point of view, using a proven design for generic infrastructure services, for persistent storage, access and management, with a bottom-up, user-centric view, based on requirements 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.

This European SCIDIP–ES Framework, whose development is coordinated by ESA and supported through the ongoing ESA Earth Observation LTDP programme,

[8]http://www.scidip-es.eu/

requires the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of the generic infrastructure services in the ES domain. Within SCIDIP-ES generic services will be tested and validated by use in Earth Science; SCIDIP-ES currently work with related ESFRI projects and other science domains to ensure generic usability of the services.

By the end of the funding period SCIDIP will have evaluated and proven the acceptance of the services across a wide range of domains as the final step towards sustainability, and have ensured a critical mass of Earth Science users. Through the definition of common preservation policies and the harmonization in the ES domain, SCIDIP-ES will moreover boost the development of the Earth Science LTDP framework 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.

This work is important because it helps our society to preserve the digitally encoded information on which we all depend, in particular those which can never be repeated, such as Earth Science measurements, and yet on which a multitude of ecological, economic and political decisions must be based in the future. The same infrastructure will allow all kinds of data to be usable by researchers from many different domains and even citizen scientists. This is consistent with the HLEG report [58] which points out that one should "Work closely with real users and build according to their requirements" and "Take advantage of growing need of integration: within and across disciplines".

The services to be provided have been identified in a large body of evidence which has been collected by PARSE.Insight 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, summarised in PARSE.Insight.

SCIDIP-ES supplements the demands from the thousands of responders to the PARSE.Insight survey with those from the ISO Audit and Certification for Trustworthy Digital Repositories [66] which is being put in place and which demands evidence touching on almost all aspects of the call. This is also the highest level of certification in the recently formed European Framework for Audit and Certification of Digital Repositories [67]. To these SCIDIP-ES adds considerations of data quality and environmental sustainability and makes advances in a number of specific areas guided by our leading edge results and our user domains.

The specific areas of the Open Archival Information System standard (OAIS, [1]) that SCIDIP-ES focuses 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 organization 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.

In order to construct the AIPs, RepInfo is needed – available from the RepInfo Registry Service and created using SCIDIP's RepInfo toolkit. The latter is an open ended collection of tools which SCIDIP-ES divides 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 SCIDIP-ES provides 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 SCIDIP-ES 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 SCIDIP-ES Orchestration Service provides a brokerage service between existing data holders and their successors. Furthermore, the SCIDIP-ES Certification Toolkit helps repositories collect evidence to submit for the ISO certification process.

There are many identifier services which claim to be persistent. SCIDIP-ES does not propose to create another such services but rather to provide an interface, the SCIDIP-ES Persistent Identifier Interface Service, to one or more chosen ones.

To support users' need to access and use data from many sources across many domains SCIDIP-ES provides a Finding Aid Toolkit to supplement the many existing domain search facilities. The services may be deployed in different ways to distribute the load and allow organizations to deploy only those services which they require.

4.4 Alliance Permanent Access to the Records of Science in Europe Network (APARSEN)[9] (Matthias Hemmje)

Digital media have become the dominant way that we create, shape and exchange information. Governments, businesses, research organisations and memory institutions, as well as individuals, have become completely dependent on digital information.

[9]http://www.alliancepermanentaccess.org/

This dependence comes with a number of major risks because of the many unresolved challenges in the long-term management, access and preservation of this information. The objective of APARSEN may be simply stated, namely to look across the excellent work in digital preservation which is carried out in Europe and to bring it together under a common vision.

The success of the currently ongoing project will be seen in the subsequent coherence and general direction of travel of research in digital preservation, with an agreed way of evaluating it and the existence of an internationally recognised Virtual Centre of Excellence. Within this broad aim, the challenges of making information accessible in the long-term require a multi-disciplinary approach across several stakeholder groups and APARSEN's first step is to bring those various communities closer together as a Network of Excellence. APARSEN states that by *"digital libraries" research we understand research intoimproving access to and use of digital resources that are held in managed collections.* Similarly by "Records of Science" in the project name APARSEN means sciences in the very broadest sense, including arts, humanities and social information.

The socio-economic impact of the knowledge generated by research activities has long been recognized due to its importance in stimulating innovation, which leads to wealth creation, growth in employment and more sustainable social development. As Commissioner Viviane Reding expressed back in 2007, …."we need to learn from each other and find together the best means to strengthen our effort in ICT research and to ensure the best use of ICT"….. "There is a need for a synchronised effort to overturn the existing inertia and drive forward growth and competitiveness".

EU innovation policy acknowledges the need for better coordination between Member States, the EC, industry and academic research communities, but action is still required to align the research for which they are collectively responsible to industrial and societal needs and expectations. The involvement of industrial stakeholders - technology providers, ICT suppliers and integrators, content providers and especially leading edge users - is vital for testing and benchmarking innovative solutions in realistic settings, but too few corporate market players have committed to being involved in driving new RTD endeavours.

This lack of industrially-led coherence is particularly significant within the digital preservation community. This spans the varied needs of the cultural, scientific and business communities as well as the vast array of public administration services. Not only are the needs fragmented by organization-type, but also by content-type - from simple rendered documents and images through to highly-structured business/government records or civil engineering records such as maintenance specifications of new nuclear facilities; from semantically-rich descriptions of content through to the preservation of scientific datasets and their experimental context. Even within a business sector like manufacturing, efforts such as the work on governance in data

preservation from the aerospace industry's LOTAR project run the risk of becoming isolated from the mainstream.

One very good example of this fragmentation is the great diversity in the use of (what should be) persistent identifier systems with different technical implementations and a huge disparity in guarantees of persistence. Sometimes this is the result of work taking place in isolation (for example, in small archives), but even large, well-funded initiatives addressing long-term problems are often resourced by short-term project-oriented mechanisms with little thought about sustainability, only the need to 'top-up' the funding with the next project. In some cases this results in good work being discontinued while, in others, extravagant claims lead to continuation of investment in industrially-irrelevant activities.

Duplicated effort due to fragmentation is potentially wasted effort. Lack of scrutiny and serious debate about the potential for innovation within ideas proposed for research funding is also potentially wasteful. Both reduce the capacity to meaningfully address the real problems of today, tomorrow and the next millennium.

Stakeholders in the emerging digital preservation market understand that digital preservation is an important Information Society issue. When digital preservation practices penetrate mainstream markets, social benefits will become visible, economies of scale will materialize, sustainability will be secured, and the impact will be multiplied across different sources and types of digital content, in larger volumes. This virtuous cycle will enable consolidation of the digital preservation market as an economic sector of the Information Society.

The next step builds on the membership of the APA, supplementing it with important players from industry and academia in order to obtain a critical mass which can consolidate the otherwise fragmented research capacity across Europe. Moreover by including the key data producers, publishers, funders, libraries, repositories and commercial interests both at national and Europe-wide level, with a huge and very varied set of users, we have the potential to exert considerable influence on the national research activities and corresponding funding.

The recent final report from the High Level Expert Group on Digital Libraries entitled Digital Libraries: Recommendations and Challenges for the Future 2 stated "A general policy framework, including sustainable custody and funding/business models, needs to be established by the key stakeholders in science and science information and national and EU policymakers. The aim is to establish the roles and responsibilities in building a European Digital Information Infrastructure that allows the access and re-use of research data and ensures their long term preservation."

APARSEN supports those aims and furthermore has close links with the EU e-Research Infrastructures through the recently formed High-Level Expert Group on Scientific Data e- Infrastructure (HLEG-SDI) which has the mandate to define a vi-

sion for 2030 and produce a detailed action plan of how to create the infrastructures needed to support the use and re-use of the very pervasive, and persistent, deluge of data with which the Information Society will have to deal. This allows APARSEN to see the direction of the future demands and therefore help to ensure the current research can provide the necessary solutions.

Additional guidance to the work in APARSEN is available from the PARSE.Insight Roadmap [2], supported by its surveys [3] and case studies with massive responses. From this one can see a consensus on the way in which the restructured research activities can lead to the future preservation service which support current re-use as well as re-use by future generations through digital preservation; this consensus seems consistent with the views of the HLEG-SDI as expressed in the final report.

5 What else needs to be taken into account? Evidence on needs in the cultural heritage sector (Nick Poole)

References

1. Digital Cultural Heritage Roadmap for Preservation, www.dch-rp.eu
2. EGI-InSPIRE project, EGI and EGI.eu: http://www.egi.eu
3. Coordination & Harmonisation of Advanced e-Infrastructures for Research and Education Data Sharing, www.chain-projet.eu