PARSE.Insight

Deliverable D2.2

Science Data Infrastructure Roadmap

Project Number	223758
Project Title	PARSE.Insight. INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Road map
Deliverable Number	D2.1
Contributing Work package	WP2: Development of a Road Map
Deliv Dissem Level	Public
Deliverable Nature	Report
Contractual Delivery Date	30 April 2008 (M3)
Actual Delivery Date	5 June 2010
Author(s)	PARSE.Insight consortium

The PARSE.Insight project is partly funded by the European Commission under the 7th Framework Programme, Research Infrastructures.

Abstract

The purpose of this document is to provide an overview and initial details of a number of specific components, both technical and non-technical, which would be needed to supplement existing and already planned infrastructures for science data. The infrastructure components presented here are aimed at bridging the gaps between islands of functionality, developed for particular purposes, often by other European projects, whether separated by discipline or time. Thus the infrastructure components are intended to play a general, unifying role in science data. While developed in the context of a European wide infrastructure, there would be great advantages for these types of infrastructure components to be available much more widely.

Keyword list

Roadmap

Contributors

Person	Role	Partner	Contribution
All partners	contributors		

Document Approval

Person	Role	Partner
David Giaretta	Co-ordinator	STFC

Distribution

Person	Role	Date	Partner
All partners			
Public distribution			

Revision History

Issue	Author	Date	Description
1.0	PARSE.Insight consortium	27 March 2009	Initial public release
2.0	PARSE.Insight consortium	18 June 2010	Include final input from Insight report

Table of Contents

1 Introduction

1.1 Purpose and scope of this document

The purpose of this document is to provide an overview and initial details of a number of specific components, both technical and non-technical, which would be needed to supplement existing and already planned infrastructures for science data. The infrastructure components presented here are aimed at bridging the gaps between islands of functionality, developed for particular purposes, often by other European projects, whether separated by discipline or time. Thus the infrastructure components are intended to play a general, unifying role in science data. While developed in the context of a European wide infrastructure, there would be great advantages for these types of infrastructure components to be available much more widely.

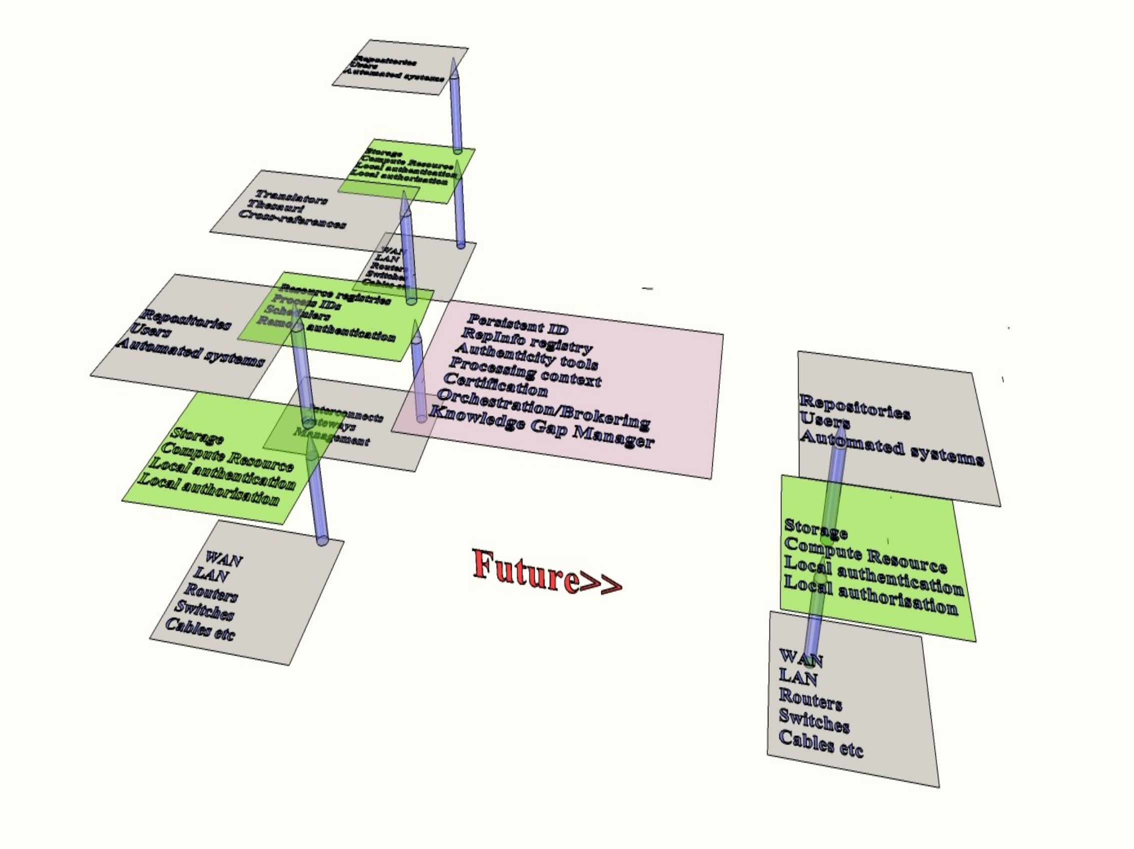


Figure 1: Infrastructure including preservation components

1.2 Science Data Infrastructure: integration with and differentiation from other infrastructures

Science Data Infrastructure is taken here to mean those things, technical, organization and financial which are usable across communities to help in the preservation, re-use and (open) access 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 (Open Archival Information System) [1], sense of maintaining the usability and understandability of a digital object . A digital object is an object composed of bit sequences.

In Europe's research landscape various actors play a role with respect to the data generated and used by the research. We have defined four main roles: funding, research, publishing, and storage/preservation. Within these four roles many stakeholders (organisations and individuals) are active with different objectives and motivations. Major influences of new developments include:

- movement to digital, but concern about digital obsolescence
- international cooperation
- new publishing models

Each community (and even on a national level) handles these transitions differently. 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 science data infrastructure must integrate with the computation and data GRID-type infrastructure [5] and provides analogous functionality in the sense of providing the linkage between islands of resources, as shown in Figure 1. The access parts of the infrastructure are provided in large part by the GRID-type infrastructure The infrastructure components provide the linkage between islands of capabilities just as the network infrastructure (e.g. GEANT (http://www.geant.net/)) links national networks and compute infrastructures (e.g. EGEE (http://www.eu-egee.org/)) link islands of compute and storage resource. The preservation aspects of the infrastructure link islands of capabilities separated by time; the re-use aspects link islands of capabilities separated by discipline and its requirements may be subsumed within those of preservation. For the former there is a one way communication from present to future and there are a number of threats which hinder the correct transmission of digitally encoded 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 (Open Archival Information System) standard, an ISO standard relating to archives, consisting of an organization of people and systems, that have accepted the responsibility to preserve information and make it available for a Designated Community.

A glossary of terms is available.

2 Demand for a Science Data Infrastructure

An associated paper summarizes the surveys which have been undertaken by PARSE.Insight and members of the Alliance for Permanent Access [2] investigating creation, re-use, preservation and publication of digital data. These surveys show a substantial demand for a science data 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 and which are pertinent to data re-use also. 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 (http://www.casparpreserves.eu/), Planets (http://www.planets-project.eu/), SHAMAN (http://www.shaman-ip.eu/), 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 preliminary survey results show that between 50% and 70% of responses indicate that all the threats are recognized as either "Important" or "Very Important", with about half supporting the need for an international preservation infrastructure. Another clear message is that researchers would like to (re-)use data from both their own and other disciplines and it is suggested that this is likely to produce more and better science. However more than 50% report that they have wished to access digital research data gathered by other researchers which turned out to be unavailable.

2.1 Quality of the evidence

The design and distribution of the surveys has emphasized comprehensiveness and wide coverage, as we believe that there is a strong need for a convincing body of evidence. There may nonetheless be some concerns about the validity of the methods and results. 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 list of threats may be either ill-founded 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 quantitative argument but nevertheless one we find very encouraging. In addition we are intending to analyse 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 that they have experienced 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 Science Data Infrastructure

We base the requirements for the preservation/re-use/access infrastructure on a broad analysis of the threats and an initial set of solutions.

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
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 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 (see [9])

4 Possible Financial Infrastructure concepts and components

It seems difficult describe an explicit business model, and indeed there may be different business models at different phases; for example one might distinguish (1) prototype (2) emerging infrastructure for early adopters and (3) a long-lived infrastructure to rely on. Certainly phase (1) would need specific funding and a number of the technical components described below have been prototyped in a variety of EU projects. For phase (2) it is difficult to avoid the conclusion that the short to medium term funding to go from prototype to stable, robust and scalable infrastructure components must be provided by the EU in the first instance, together perhaps with major stakeholders such as the members of the Alliance for Permanent Access [2]. The longer term business model needed for phase (3) must clearly be linked to the business models for the rest of the infrastructure on which the components described here depend, for example the basic network.

It is worth making a number of observations, for example 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 necessarily suffer the problem of scaling with the amount of information being preserved. For example one piece of Representation Information may be used to describe 1 billion data objects.

A 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 linear cost relationship. Nevertheless such a facility may be able to supply added value services such as disaster recovery and integrity checking.

Cost/benefit analyses are likely to be very highly instance specific yet some common models are essential if judgments are to be made about what can be afforded. A common framework for at least collecting the information would be useful if a good understanding of the important parameters is to be gained.

The Blue Ribbon Task Force report [5] provides a number of important considerations including the important idea of buying future options. Here funders put off making a long term committment, instead funding preservation for a sufficient time that allows future decision makers to be in a position to make that decision.

Another issue concerns the incentives for data producers to make their data available; for example funders could require, perhaps by making part of the funding depend upon it, that steps are taken by researchers that their research outputs are make available for preservation.

There are a number of resource generating value-added services which can be associated with digital preservation. These include services such as Portico [6], used by some libraries for an outsourced preservation service, and the related services such as the ones prototyped by PEPRS [7] which keeps track of which electronic journals have arrangements in place for digital preservation so that journal articles can be accessed over the long term. However funding for such services derives from the usual sources –

national research funders, EU or research organisations. New sources of funding may come from advertising revenue where there is wide public interest in the data, for example images of the Earth or the sky.

5 Possible Organisational and Social Infrastructure concepts and components

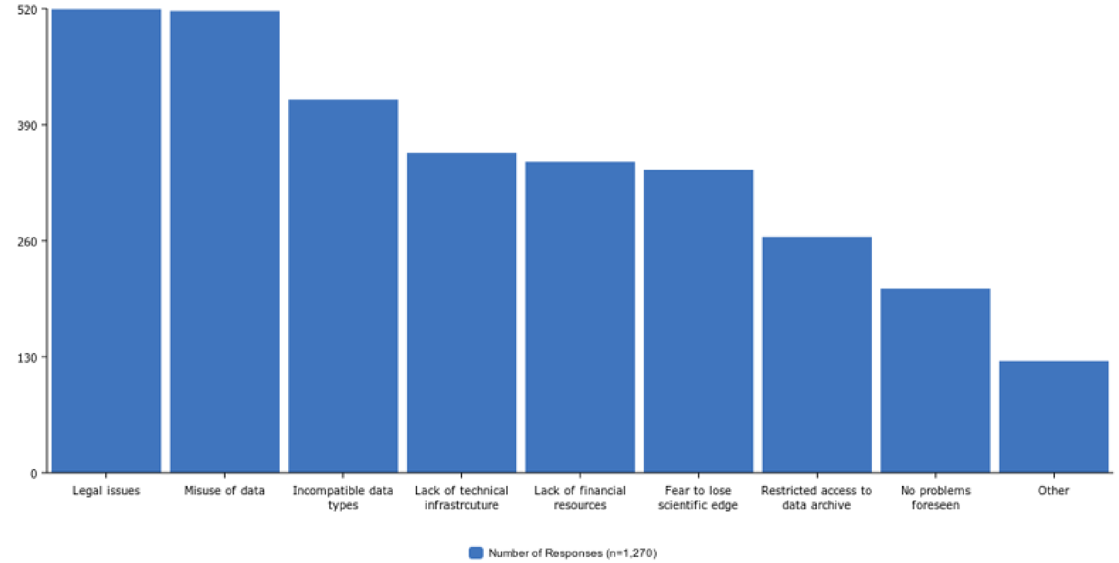
It is clear that a number of the infrastructure components described in the previous sections are themselves archives which need to preserve digital information over the long term and which therefore themselves require the support of that very preservation infrastructure. For example, the threat of *The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future* should be taken into account by securing the handover to another host organisation. Also, 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, although commitment to provide a service for an indefinite time tends not to be popular. An alternative approach is to move the responsibility to an arms-length or consortium-based organisational structure, such as the Alliance for Permanent Access [2]. This structure is bringing together key stakeholders in many sectors and may play a key role. Even this may need to be underpinned by governmental guarantee in order to provide real confidence in the infrastructure's longevity.

Aside from the organisational component, social/behavioural aspects must be considered. For example, a science data research infrastructure must facilitate data sharing and data mining. However researchers do have concerns about this; indeed, it has been (jokingly) said that data sharing/mining means either *"this data is* ***mine*** *[and no one else's]"* or else *"my data is mine, and now your data is mine [to use as I like]"*

More light can be shed on this through the survey results conducted by PARSE.insight. While a majority of researchers say they would like to make use of the research data of others, the researcher's survey also shows that a considerable number of researchers foresee problems in making their own research data available for others. No more than 25 % make their data available for everyone (against close to 60 % who share it within their research group). What are the problems mentioned? Over 40 % are afraid of misuse, around 40 % foresee legal problems (e.g. breach of privacy, misuse of anonymous surveys, etc), between 25 and 30 % mention technical problems (lack of infrastructure, incompatible data, access restricted, etc).

RESEARCH - Do you experience or foresee any of the following problems in sharing you data? (multiple answers possible)

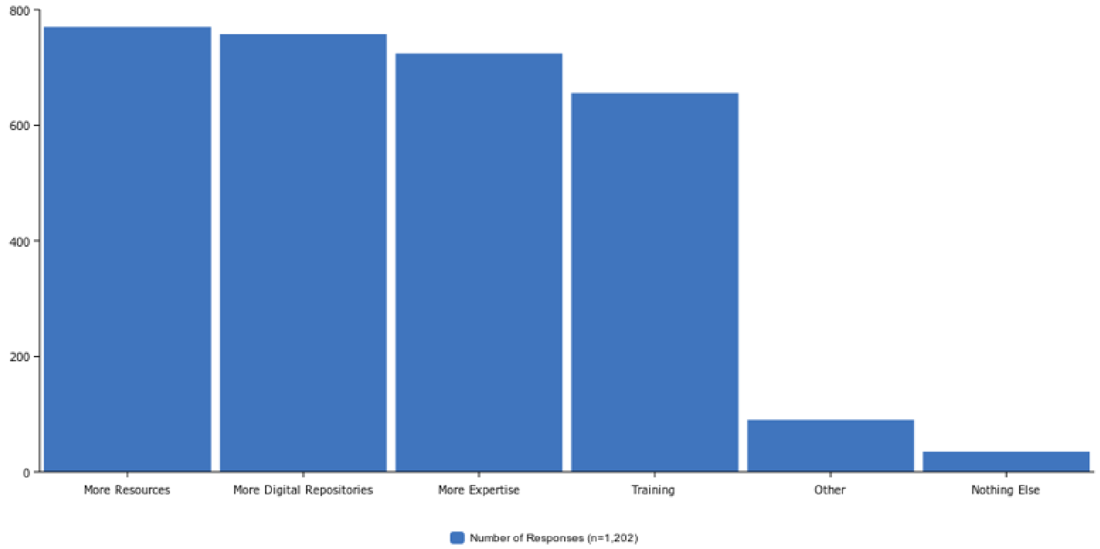


Source: RESEARCH - Do you experience or foresee any of the following problems in sharing you data? (multiple answers possible) on Swivel

[1]

This implies that even when a technical infrastructure is in place for the preservation of research data, the current behaviour patterns may prevent people from using it. This observation is enforced by another outcome of the survey regarding additional needs to operate in a digital research environment. Many researchers noted they feel the need for more resources (64 %), more digital repositories (63 %), more expertise (60 %) and training in how to preserve and share your information (54 %). In addition, they noted that guidelines/manuals on preservation, workshops, a knowledge platform and user-oriented training sessions are very important.

RESEARCH - Apart from an infrastructure, what do you think is needed to guarantee that valuable digital research data is preserved for access and use in the future? (multiple answers possible)

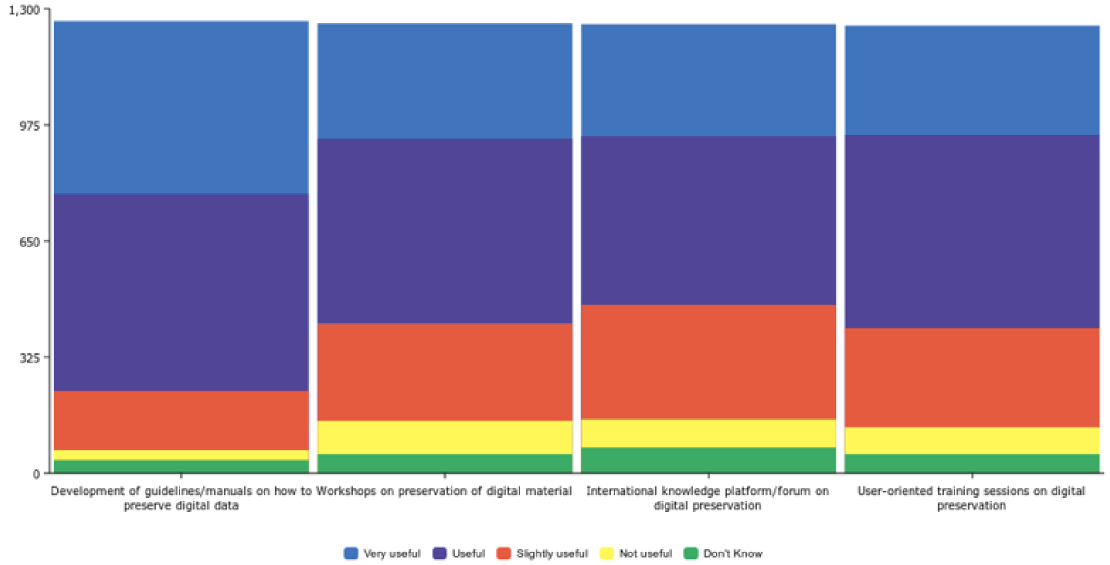


Source: RESEARCH - Apart from an infrastructure, what do you think is needed to guarantee that valuable digital research data is preserved for access and use in the future? (multiple answers possible) on Swivel

[2]

---

[1] https://www.swivel.com/charts/9114-RESEARCH-Do-you-experience-or-foresee-any-of-the-following-problems-in-sharing-you-data-multiple-answers-possible-



RESEARCH - Do you think the following initiatives would be useful for raising the level of knowledge about preservation of digital research data?

Source: RESEARCH - Do you think the following initiatives would be useful for raising the level of knowledge about preservation of digital research data? on Swivel

[3]

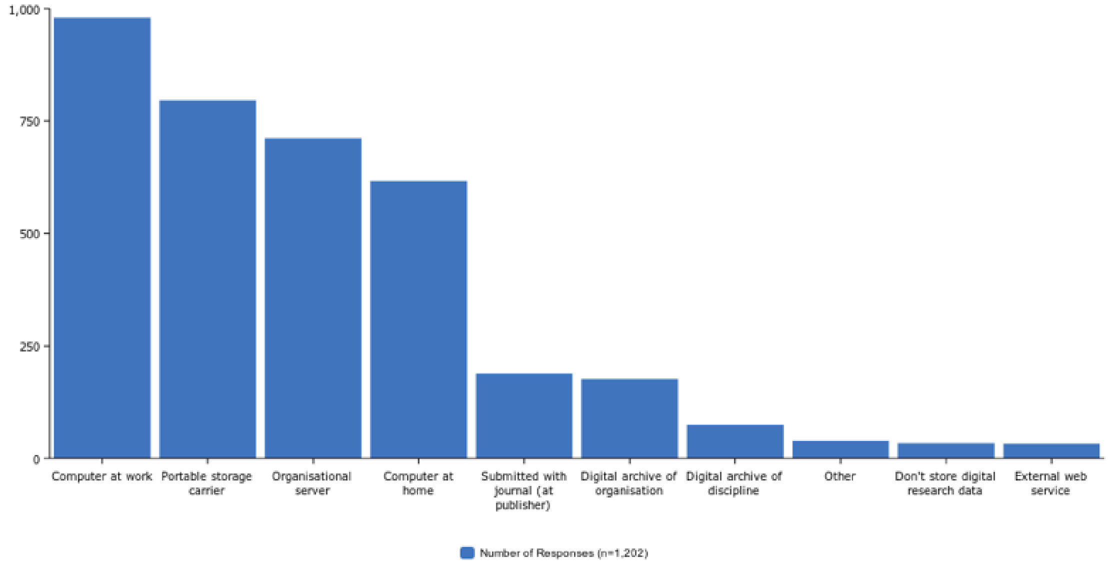
Although large scale facilities often have archived copies (backups, held for at least a little while) of the data they are used to create, the data created by individual researchers are often treated less well. Institutional repositories have not been great magnets for such data. Instead, most researchers that responded on the survey still use their own computer at work to store their research data (81 %).

---

[2] https://www.swivel.com/charts/9126-RESEARCH-Apart-from-an-infrastructure-what-do-you-think-is-needed-to-guarantee-that-valuable-digital-research-data-is-preserved-for-access-and-use-in-the-future-multiple-answers-possible-

[3] https://www.swivel.com/charts/9139-RESEARCH-Do-you-think-the-following-initiatives-would-be-useful-for-raising-the-level-of-knowledge-about-preservation-of-digital-research-data-

RESEARCH - How do you presently store your digital research data for future access and use, if at all? (multiple answers possible)



Source: RESEARCH - How do you presently store your digital research data for future access and use, if at all? (multiple answers possible) on Swivel

[4]

To encourage and facilitate the behavioural changes PARSE.insight identified a number of benefits for researchers to get their data preserved and re-used:

- Safe your data for future research: 53 % of the researchers that responded on the survey mentioned that they have experienced data created previously by other researchers was not available (anymore). By offering a guarantee that data deposited in a digital archive is kept for the long term, researchers do not have to worry about taking archiving actions by themselves.
- Citability: with persistent identification, preserved data can be found and re-used any time and any place. Based on the science data infrastructure, data sets could be cited as well. This way, a researcher gains extra visibility and can receive credits for publishing the data.

This requires several organisational components to be put in place:

1. Policies: in some countries mandates exist for depositing research data and in some cases funding agencies require so. But clearly, this shall not be enough as certainly not all researchers seem to obey the mandate.
2. Robust and reliable deposit places, where researchers can be sure their data will not get lost, be corrupted or misused. Reliable also means with the right access mechanisms, perhaps even some kind of access permission system for retrieval via the creator of the data. In response to being asked where they would like to store their data, the three best scoring options are: digital archive of their institute (63 %), discipline based archive (60 %) and at the publisher (47 %).
3. Elements that increase comfort levels so that new users will know how to use and interpret the available data. And that new users will not take these data out of context. This could be achieved by a good linking system between the

---

[4] https://www.swivel.com/charts/9128-RESEARCH-How-do-you-presently-store-your-digital-research-data-for-future-access-and-use-if-at-all-multiple-answers-possible-

data and all publications that exist for and mention these data. In the survey, some 96 % of respondents say they publish about their data in journals of publishers – surely these articles will contain a section on methods and protocols where new users can find how the data were gathered, if there are any restrictions on how to (re)use them and what the context of these research data is.

4 Communication and awareness around these issues.

5 Have publication of data as valued and referenced as is a publication of a paper in a journal.

These incentives can be grouped in soft and hard encouragements. Furthermore, each stakeholder in research can play a specific role in this. Below, a table is depicted which outlines the *carrots and sticks* that can be applied by a stakeholder.

Stakeholder	Can offer carrots	Can offer sticks	Has dependencies	Has boundary conditions
Data manager	Training support, guarantee trustworthiness	Archiving regulations	Funding, standards and guidelines	Institutional policy, infrastructure
Publisher	Citability, Accessibility, High quality papers	Journal rejects if data not supplied and preserved or if DOI for data is lacking	Good infrastructure, Persistent identifiers for linking	Editors agree, reviewers can review, agreed standards on how to review data
Funder	Extra funding for preservation of research data in trusted digital archive. Extra funding for sharing data with other researchers and communities	No funding and no payment if data not deposited, require proposal to address preservation aspects	Budgets	Manageable, controllable ? Is infrastructure in place ?
Researcher	Reproduceable results, basis for more research, citability and recognition	Embargoes, no access for researchers who do not make data available	Good infrastructure	Part of bigger group
Research Institute	High quality research environment, no worry about cost or safe deposit places,	Rules of the house, limit budgets if not done Restricting	Budgets, good infrastructure	Common standards

	easy tools for automatic generation of contextual information for data	budgets		
Employer	Eternal fame	Labour condition	Common practice elsewhere	Labour laws
Policy makers	More budgets, awareness raising	No budgets	Good infrastructure	Common practice
Research Peers	Help improve results, add citations, crowd sourcing, become part of social network (online collaboratory)	Community pressure, punish by means of oblivion	Infrastructure, common practice	Common practice
Library	Accessibility and retrievability, ensure trustworthiness of digital archive	Oblivion	Persistent identifiers, good linking between publications and data, trusted digital repository, policy, standards and guidelines	Common practice
Data collectors	Preservation requirements integrated into way data is collected	Data will be lost or never found	Persist Identifiers, good metadata schemes, central registries like TIB and Datacite	Common practice

5.1 Scenarios

***SCENARIO**: the researcher*

After several years of hard work and gathering tremendous amounts of raw data, a researcher gets her paper accepted for top-journal AAA – the ground breaking conclusions of her research stem from a new analysis methodology that she developed herself to allow for proper processing of the data. Her paper extensively describes the data set as well as the methodology, in terms of its richness as well as its shortcomings. She decides to make the data available for sharing with other researchers via the subject specific data repository at an international research institute (or at a national data centre) and adds the software code of the analysis methodology to it. But she has the following demands:

- *she wants anyone looking at the data or at the software to first consult her extensive research paper in journal AAA, to avoid misuse of the data or tak-*

ing conclusions out off context;

- *she wants any researcher wanting to re-use these data to be aware of the special methodology needed for the proper analysis and how that was applied in the software (hence: to read the paper);*
- *she wants her dataset to be properly cited when re-used, with a pertinent and persistent link to where the dataset resides, idem for the software code and of course with a link to her original research paper in journal AAA;*
- *she wants the data to be available as a separately citable item, counted in all significant citation scores;*
- *she wants to be sure that any reader of her research paper in journal AAA can easily link to the datasets and the related software.*

SCENARIO*: the library*

A National library in Europe has changed its policy to broaden its scope of their digital collection. From now on, data sets and accompanied tools for analysis of that data (software) in the field of Social Science and Humanities are part of their digital collection plan. In conjunction with that, they have adjusted their existing digital archive to also cope with digital material other than electronic scientific papers, book and magazines. To do so, they put in place extra preservation strategies that deal with this new type of content such as an emulation delivery service for running research software. Furthermore, they enriched their catalogue functionality with a linking mechanism for data sets and improved Digital Rights Management to allow several levels of access such as "open access all time", "open access after X period", "restricted access - only after authors approval". However, the descriptive information of the data will always be open.

With the new functionality in place, the National library is technically speaking ready for accepting data and software. To ensure researchers and their institutes are aware of this, they started a audit and certification procedure so that they can proof that their organisation is a trusted place for research output. In parallel, the library contacted several research organisations in their domain to set up archiving agreements. Part of these agreements is that researchers should spent a minimum amount of their valuable time to selection and archiving research data. The library committed themselves to this by building flexible tools for selection and metadata capture in cooperation with the research organisations. Furthermore, they will offer training courses of two days for each researcher that will start using the archive. In turn, each research institute pays a yearly fee for depositing the research output at the library. Also, the institutes are intended to form an international group of experts that will support the developments of data standards and selection procedures in their domain.

***SCENARIO**: the* the funding organization

As part of a significant new research initiative, a national funding organisation in Europe issues a call for research proposals in the area of behavioural social sciences. The theme is about eating habits of adolescents, to be measured and collected and analysed over a period of 25 years, with reporting on trends every 5 years. The call is hugely oversubscribed but the multitude of proposals deliver a wide spectrum with some very interesting different approaches. The aspects to compare and analyse cover many different areas and topics – all of them innovative and promising in their own right. A substantial part of the research costs involve the collection of large amounts of data. Given the size of the grant and without further interference, a maximum of 2 or 3 proposals could be granted.

In order to allow more output for the same research budget, the funding organisation decides to concentrate data collection within one research group and to require them to undertake it in such a way that several other groups who submitted good proposals can use the same data. It means that one group will need more money, but others substantially less, with the positive effect that between 8 and 10 proposals could be granted! The funding organisation starts an inventory of what is required in such a collaborative set up, based on the sharing of data. Enthusiastic about this idea, the Funding Organisation decides to require any data collection under this call to be stored and preserved properly so that also in the future the data can be accessed and re-used for new initiatives.

***SCENARIO**: privacy aspects*

A research group is about to publish a study in the well regarded journal Science&Nature. The study has analysed and compared the data of the first 5 years of 6 million EMR's (electronic medical records) and drawn shocking conclusions about:

- *Indications for many untreated chronic diseases in early stages;*
- *Incorrect diagnoses in more than 20 % of cases by 60 % of general practitioners;*
- *Prescription of wrong medicines in 32 % of cases where medicines were prescribed.*

The journal Science&Nature requires since early 2008 that any research of which the conclusions are based on datasets, make these datasets available for peer review but also for those willing to dig into it. But in this case there is a serious problem with privacy aspects as many data are retrace-able to individual patients. There is also the ethical aspect to the study: the indications for early chronical diseases: should the patients to which this applies be informed or not?

The journal Science&Nature is prepared to limit the requirement for availability of data to the peer reviewers only who shall look at the datasets only from the point of the way the data was gathered and analysed – without entering specific EMR's. With reference *to their normal preconditions that data can only be made available to readers if no security, privacy, or other commercial or ethical aspects are in the way, the journal shall drop its requirement to make the data available.*

SCENARIO*: training those who (will) deposit research data*

DataManager Z works at a national research institute that was one of the very first to require from its researchers that all data collected for their research is stored in its datacentre. It took a little while before the practice had found its pattern, but now Mr Z gets inundated by datasets, of any sort, size and shape. On the website of his datacentre a manual was published about the formats to use, the metadata to add and the descriptors to use to ensure optimal storage, preservation and re-use of the data. But to no avail. The biggest problem is that hardly any of the researchers have been thinking through what later storage, preservation and re-use implies for the way the data is collected. And after the fact, it is difficult to repair any flaws, if only because many of the researchers have already moved on to their next project.

Z decides that the situation will not get any better if the researchers (most of them young and enthusiastic but relatrively unexperienced PhD? *'s) do not get any training. So he starts one-day training sessions, widely announced, but with very few attendants. Then he thinks of e-learning modules. New researchers who join the institute are required to work through the module in their first month and finish it with an online test. Also, any research proposal at the institute requires a separate paragraph about the data collection, preservation and possibilities for re-use.*

5.2 Next steps

- Make organisations aware that digital preservation is not only a technical challenge but also requires adjustments to their policies and procedures.
- Define and apply standards for exchange of data sets across research institutes and archives.
- Define and apply standards for openness of data (aka Creative Commons for data).
- Agree on a checklist for digital archives to become "trusted".
- Develop training courses to teach researchers how to cope with digital data (awareness raising).
- Develop e-learning modules for training researchers to work with data sets and how to archive and share them.
- Develop guidelines for researchers and their institutes to come to common practice data formats which are suitable for archiving.
- Build a cross-domain virtual platform for researchers to learn about best practices in sharing and archiving of data.

- Commission expert panels (for each discipline) to support the selection process of what needs to be preserved and what not.
- Develop easy-to-use tools for data selection and archiving preparations (limiting misinterpretation of data and fragility of data formats).
- Demonstrate citability of data sets within and across disciplines.

5.3 Final destination

- Safely deposited data sets at a (perhaps limited) number of trusted digital archives.
- Updated policies and procedures of research institutes taking long-term preservation of their assets into account.
- Data sets (or access copies) that can be cited amongst disciplines and for which researchers can be credited.
- Respected access mechanisms to protect data and researcher from misuse and misinterpretation.
- Trained researchers that are aware of digital fragility and how to cope with that (selection, Representation Information including file formats and data semantics, descriptive information).

5.4 Relevant projects, policies, organisations, activities:

- ISO Repository Audit and Certification work [9]
- DANS Data Seal of Approval
- OAIS reference model
- Creative Commons
- Alliance for Permanent Access

6 Possible Policy infrastructure concepts and components

There are a number of broad policies or statements of intents about preservation, re-use and (open) access. Although it is not clear when or whether these will converge, it is clear that there will almost certainly be a variety of such policies for the foreseeable future. The preservation infrastructure must be able to operate in this environment. Nevertheless alignment of policies will undoubtedly make the task simpler, for which co-ordination at national, EU and international levels, including EU and transnational consortia of key stakeholders such as the Alliance for Permanent Access [2], would be essential. Clarity is also needed in policies which not just encourage but also require researchers, perhaps with financial incentives, to deposit their data, and which also indicate practical ways for them to do so.

It is important that policy makers distinguish between **open access** on one hand and **digital preservation** on the other. Those whose main interest is open access – if it is to be maintained over the long term - should understand that it is equally dependent on digital preservation, as is non-open access.

6.1 Deployment and Adoption

The need for an infrastructure on an international scale is evident. To ensure that such an infrastructure will be supported by all stakeholders across Europe and beyond, a well-defined strategy is needed to stimulate its adoption. This strategy can be considered from two perspectives: a bottom-up view, representing the view of the end-users (researchers, publishers, data managers, etc) and a top-down view which represents the perspective of the initiators of the infrastructure.

The bottom-up perspective currently gives a view on many initiatives taken on sharing data amongst researchers within their research domain as mentioned in the previous section (e.g. GEANT, EGEE). These national or domain-specific solutions (islands of capabilities) are mostly developed to enable interoperability between different science stakeholders. The clustering of information resources is an ongoing process already and will eventually lead to larger networks that allow stakeholders to share information. However cross-domain cooperation will still be limited due to incompatibility of these domain-specific infrastructures. The solutions often do not share a standardised and certified approach, which limits overall sustainability of the infrastructure. While respecting the existing solutions, it is a challenge to achieve a global infrastructure that not only allows researchers to share data, but also to keep the information trusted, reliable and secure.

To achieve better sustainability and interoperability, the top-down approach can help by promoting the foundation of guidelines and recommendations for sustainable data archives. The Repository Audit and Certification work [9] aims in this direction. Moreover, standards should be promoted which are compliant with a trans-national infrastructure, but also are easy to adopt in the already existing networked domains. The EU as well as other international bodies can play an important role in this process.

It is worth noting that Audit and Certification can apply not just to the data holdings but also to most of the other infrastructural components which are mentioned in the next section. These components depend upon their holdings of information, for example Representation Information, which must itself be preserved over the long term. Thus those **infrastructure components should themselves be audited and certified** if the infrastructure itself – or at least the information on which it depends – is to be usable over the long term.

The benefit of this top-down approach not only ends with better interoperable and sustainable networks, it also draws a clear scenery of the European science landscape, allowing new stakeholders to build a business model on top of the infrastructure. Researchers are assured that their data is compatible and safe because of certification and legislation while new businesses can offer new services on top of this secure layer of the infrastructure.

A good example is the OAIS Reference Model (ISO 14721:2003), which has become a worldwide adopted standard for building a sustainable digital archive. Today, various vendors developed their own archiving solutions and bring them to the market.

7 Virtualisation of Policies, Resources and Processes

Virtualisation is a commonly used technique in systems to insulate services from underlying implementations. The science data infrastructure described here is implemented by services including management, trust, workflow, data storage and other resources. In order to insulate the science data infrastructure components from changes it is necessary to try to virtualise access and use of all these. Virtualization would for example facilitate the migration between preservation environments, i.e. enabling policy enforcement across systems.

SCENARIO
Due to its size, a large scientific dataset has to be stored across multiple distributed locations. These storage locations are maintained by different organisations using diverse hardware/software infrastructures. Researchers who wish to access the dataset are provided with a uniform interface, hence they do not need to be aware of the actual physical location of the data. Data managers are provided with a standardized set of actions, which are then mapped to concrete operations and executed by the respective underlying infrastructures. Computing-intensive operations such as format migrations might be scheduled and submitted to external (grid-based) services.

Next steps:

- Specify standards promoting the interoperability between services, grid operations and existing archive systems.
- Scalable storage abstractions capable of handling increased data volume without impacting the running of the archive
- Support for data replication to geographically disparate storage resources.
- Provision of logical namespaces for resources, data and users.
- Define data virtualisations for common data objects

Final destination

- Infrastructure independence, collections can be moved across preservation systems without any loss of information.
- Management virtualization, seamless federation of preservation environments while maintaining control over policies, processes and resources.

Relevant projects, policies, organisations, activities:

- SHAMAN (http://www.shaman-ip.eu/), Chronopolis (http://chronopolis.sdsc.edu/mediawiki/index.php/Main_Page), CASPAR (http://www.casparpreserves.eu/), iRODS (http://www.irods.org)

8 Technical Science Data concepts and components

Each of the solutions is analysed next. For each solution there is a particular need to review the existing digital preservation projects, review the proposals and identify open issues. The Warwick workshop report [10] is also relevant here.

8.1 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.

SCENARIO

A dataset created by one researcher may need to be used by a second, either contemporaneously or at some later time. This second researcher may come from a different discipline and use different analysis tools. In order to avoid producing misleading results he/she must be able to understand what the data actually means. For example, given an astronomical image in the current FITS format, with its several variants, the researcher would need to be able to extract the values of the pixels of the image from what may be quite a complex and highly tailored digital object. In order to use an analysis tool one would need to know how to deal with these pixel values, their units, their coordinates on the sky and the way in which the photons have been selected e.g. the bandpass of the filters used. Representation Information is the OAIS term for everything that is needed in order to understand a digital object. A registry would help to ensure that the required Representation Information is available in the future and across disciplines.

Next steps:

- Representation Information Registry holding copies of Representation Information of all types which can be shared and enhanced by contributions from many people.
- Virtualisation techniques to facilitate easier integration into contemporary tools
- Preservation features should be embedded in the "creation" environment, automating/facilitating the generation of necessary representation information (data, models, assumptions, configurations, ...).
- 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. Automated capturing of the creation and processing context.

Relevant projects, policies, organisations, activities:

- CASPAR (http://www.casparpreserves.eu/), Planets(http://www.planets-project.eu/), DCC (http://www.dcc.ac.uk/), JISC (http://www.jisc.ac.uk/), OAIS (http://public.ccsds.org/publications/archive/650x0b1.pdf), SHAMAN (http://www.shaman-ip.eu/), nestor (http://www.langzeitarchivierung.de/)

8.2 Sharing of information about hardware and software

Ability to share information about the availability of hardware and software and their replacements/substitutes:

SCENARIO
A performing artist finds a masterpiece of (formerly) modern music which requires a signal processing system which used to run on an Apple MacIntosh to add a special type of reverberation to the sound. The artist has a number of options including finding the signal processing software together with a working Apple MacIntosh, or an emulator running on his/her computer.
A way to sharing information about hardware and software would facilitate the re-performance of this masterpiece.

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 (http://www.casparpreserves.eu/), KEEP (with regard to emulation) (http://www.keep-project.eu), Planets (Ada asks: should this be PlaneTs), nestor (http://www.langzeitarchivierung.de/)
- Need for a software archive

8.3 Authenticity of a digital object

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 (http://cidoc.ics.forth.gr/), OPM (http://openprovenance.org/). 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.

SCENARIO
A virtual reconstruction of the Taj Mahal created at the start of the 21st century shows that there have, 50 years later, been subtle damage caused by a local development. The developer disputes this and argues that the digital data on which the virtual reconstruction has been made is not what is claimed. What evidence can and should be provided to support the claims of authenticity and hence save the Taj Mahal? Strong techniques and support tools are needed to allow curators to support claims of authenticity

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
- Develop tools to map provenance to authenticity tools
- Maintain the chain of evidence through (automated) digital audit (provenance) trails by embedding support for capturing knowledge about the actual operations performed

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 (http://www.casparpreserves.eu/), SHAMAN (http://www.shaman-ip.eu/), nestor (http://www.langzeitarchivierung.de/)

8.4 Digital Rights

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. There are several digital rights expression languages in the academic community and commercial world - some are being standardised – the infrastructure must be able to cope with this variety and their evolution and possibly of the underlying rights. 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.

SCENARIO
A piece of software was produced by an inventor and is protected by a user key which must be renewed every year. Several years after the death of the inventor the software is needed by a researcher in another country with a different legal system. What restrictions on usage are there under this rather different system? Even if the software could legally be used, how can the appropriate software key be created? A way is needed to be able to handle the link between the rights and restrictions originally associated with the digital object and the legal system under which it is eventually used.

Next steps:

- Share information on how constraints, which DRM (Digital Rights Management) 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 (http://www.casparpreserves.eu/), ARROW (Accessible registries of rights information and orphan works towards Europeana) (http://www.arrow-net.eu/), nestor (http://www.langzeitarchivierung.de/), KoLaWiss (http://kolawiss.uni-goettingen.de)

8.5 Persistent Identifiers

Need 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.

SCENARIO

A researcher reads a paper in a journal which refers to a dataset which he realises can be re-analysed and combined with some new data he has recently obtained. The paper has an identifier string for the dataset which after some investigation he sees is some sort of a "persistent identifier". Unfortunately the originator of that system is long gone, the DNS entry for the identifier name resolver system host has lapsed and the database system which was used is not available.

A more permanent persistent identifier system is needed which itself has the appropriate longevity with committed long-term financial and social support.

Next steps:

- Review the existing persistent identifier systems and their technical, organisation and social underpinnings with respect to longevity and scalability
- Develop or adopt a sufficiently scalable/maintainable identifier system
- Investigate potential organisational underpinnings and the links to, for example, the EU or USA.

Final destination

- An identifier system for locating and cross-referencing 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 (http://www.doi.org/), DNS (Domain Name System), CASPAR (http://www.casparpreserves.eu/), URN (Uniform Resource Name), nestor catalogue of criteria for trusted PI-systems (http://www.langzeitarchivierung.de/), XRI (Extensible Resource Identifier), DPE (http://www.digitalpreservationeurope.eu/)

8.6 Transfer of custody and brokering services

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.

SCENARIO
An archive finds that its funding agency has been wound-up and the archive must close in six months time. Moreover the data holdings are currently in a set of inter-related database tables with embedded binary large objects, and a sophisticated access system with much embedded business logic. How can the archive find someone willing to look after its holdings and how can they be handed over in practice?
Although individual repositories tend to have specialised access systems tailored to help their users, attention must also be paid to ensuring that the holdings can be handed over if/when necessary, and appropriate tools and techniques are needed to help do this.

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 (http://www.casparpreserves.eu/), SHAMAN (http://www.shaman-ip.eu/, OAIS http://public.ccsds.org/publications/archive/650x0b1.pdf)

8.7 Certified repositories

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

Although one cannot guarantee anything into the indefinite future there has, 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.

SCENARIO
A funding agency wishes to instruct its researchers to deposit their data into one or other of the long term archives it will support. This will involve a large and continuing commitment of resources. How can the funder be sure that the archives it wishes to support are up to the job?
An internationally recognised certification system would give funders and depositors a

way to distinguish and evaluate archives.

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), DCC (http://www.dcc.ac.uk/), DRAMBORA (http://www.repositoryaudit.eu/), OAIS (http://public.ccsds.org/publications/archive/650x0b1.pdf), Alliance for Permanent Access [2], EU (http://europa.eu/), NSF (http://www.nsf.gov/), JISC (http://www.jisc.ac.uk/), nestor (http://www.langzeitarchivierung.de/)

9 Aspects excluded from this Roadmap

A number of science data related activities have been excluded from this document on the basis that (1) they provide the islands of capabilities and therefore (by definition) are not infrastructure and (2) it is not at all clear that an infrastructure can be created to support these activities, however this must be reviewed. Access methods have not been discussed above because they are expected to be largely provided by GRID-type capabilities, although clearly infrastructure such as persistent identifiers will play an important role in access services.

The list of excluded topics is as follows:

- Specific organisational budgets
- Decisions of what to preserve i.e. appraisal – although clearly some co-ordination would be useful
- Specific domain software
- Specific national legal aspects – although the ability to cope with a variety of these must be built into the infrastructure.

References

[1] Reference Model for an Open Archival System (ISO 14721:2002), 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

[2] Alliance for Permanent Access http://www.alliancepermanentaccess.org/

[3] Report on Roadmapping of Large Research Infrastructures, 2008, OECD International Scientific Co-operation (Global Science Forum). Retrieved from from http://www.oecd.org/dataoecd/49/36/41929340.pdf

[4] Understanding Infrastructure: Dynamics, Tensions and Design http://www.si.umich.edu/~pne/PDF/ui.pdf

[5] http://www.informit.com/articles/article.aspx?p=169508&seqNum=5

[6] Portico, http://www.portico.org/digital-preservation/

[7] Pilot E-journals Preservation Registry Service, http://edina.ac.uk/projects/peprs/

[8] NSF Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Web site http://brtf.sdsc.edu/

[9] Repository Audit and Certification ISO working group – see http://wiki.digitalrepositoryauditandcertification.org/

[10] Warwick workshop report http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf