Components for a Science Data Infrastructure – preservation and re-use of data

David Giaretta
STFC, Rutherford Appleton Laboratory, Didcot, Oxon OX11 0QX, UK
Email: david.giaretta@stfc.ac.uk

ABSTRACT

There are a great number of different organisations and projects creating, using, preserving and re-using data. It seems that each discipline and each organisation claims specialised needs and therefore dedicated funding, Nevertheless this paper will show evidence captured by the PARSE.Insight project that the concerns of data producers, users, publishers and managers, about preserving and (re-)using data, have great similarities. An analysis of these common requirements allows one to identify a Roadmap for components of an infrastructure to support preservation and use across a wide variety of disciplines and data types.

Furthermore these components may be seen to supplement the layers of capability provided by the network and the GRID – normally illustrated by various layers. Essentially at the lowest level there are isolated islands of capability, with for example cables, routers etc; these are joined by interconnects and gateways. On top of these each organisation has storage and compute capabilities; components such as resource registries, schedulers and process controls allow these to interoperate. A third layer embodies the repositories; components such as cross-reference services allow these to work together.

As we move to the future we envisage the need for another set of components to assist us in faithfully communicating our current information (not just the bits!) to future users. Because we define, following OAIS, preservation in terms of use, these same components help use and re-use of, and adding value to, contemporaneous data. This paper will provide details of the Roadmap and an initial analysis of how the gap between what is required by this Roadmap and what is already planned as infrastructure.

INTRODUCTION

PARSE.Insight (http://www.parse-insight.eu) is a two-year project co-funded by the European Union under the Seventh Framework Programme. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research. The problem is how to safeguard this valuable digital material over time, to ensure that it is accessible, usable and understandable in future. The rapid pace of change in information technology threatens media, file formats and software with obsolescence, and changing concepts and terminology also mean that, even if data can be read, it might not be correctly interpreted by future generations.

Many initiatives are already under way in this area, and the aim of the PARSE.Insight project is to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The project will conduct surveys and in-depth case studies of different scientific disciplines and stakeholders and will base its results on these findings, as well as knowledge of ongoing developments.

PARSE.Insight is closely linked to the Alliance for Permanent Access to the Records of Science (http://www.alliancepermanentaccess.org). The output from the project is intended to guide the European Commission's strategy about the science data research infrastructure.

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 sense of maintaining the usability and understandability of a digital object .

In Europe’s research landscape there are a number of recognisable roles played by various actors. The PARSE.Insight project has 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 and provides analogous functionality in the sense of providing the linkage between islands of resources, as shown in Figure 5. 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) links national networks and compute infrastructures (e.g. EGEE) 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.

INFRASTRUCTURE

The purpose of science data infrastructures is to enhance the overall productivity of European science. Although the mechanism for monitoring increased productivity is not straightforward, nevertheless surveys show that the majority of researchers do believe that access to other people’s data will improve their scientific efficiency and furthermore promote multidisciplinary activities which have so far been difficult to perform.

The Demands on a Science Data Infrastructure

Surveys have been undertaken by PARSE.Insight project investigating creation, re-use, preservation and publication of digital data. These surveys, which, with the associated case studies, have several **thousand** responses from across disciplines and across the world, show a substantial demand for a science data infrastructure.

The full analysis of the results will be available on the project web site (http://www.parse-insight.eu). 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.

The major threats are as follows – expressed in non-technical language:

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

Another clear message is that researchers would like to (re-)use data from both their own and other disciplines and 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.

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

SOLUTIONS TO THE THREATS

Solutions to the threats can be described, in more appropriate terms, described next, using terminology from the Reference Model for an Open Archival Information System (OAIS), and the CASPAR project (http://www.casparpreserves.eu) .

Fundamental models and workflows of digital preservation

CASPAR follows the OAIS Reference Model concepts and terminology, extending them where OAIS does not provide enough detail. OAIS contains a number of models. The most important of these is the Information Model, shown in Figure 1.

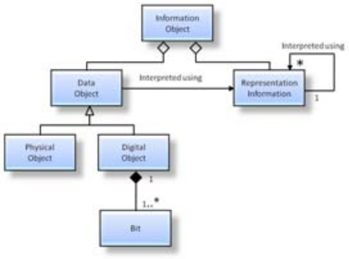

Figure 1 OAIS Information Model

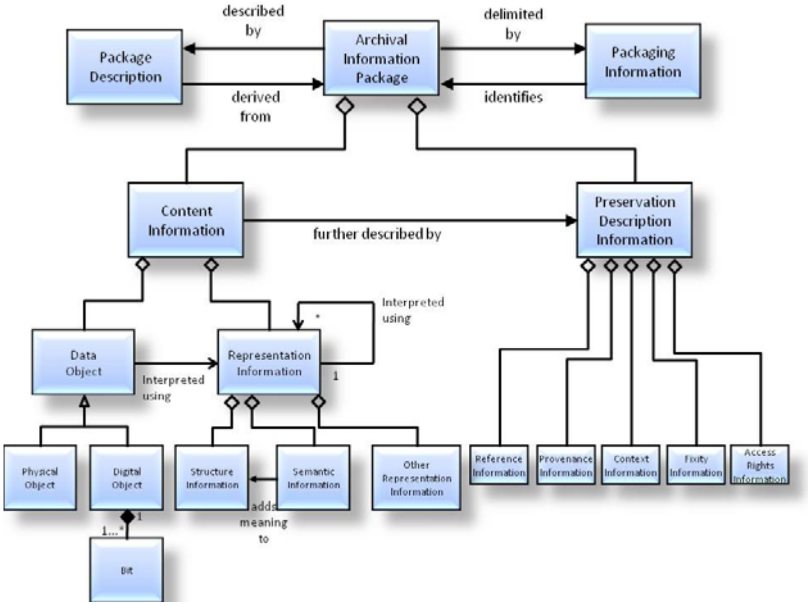

Figure 2 OAIS Archival Information Package (AIP)

Representation Information

The UML diagram (Figure 1) means that

- an Information Object is made up of a Data Object and Representation Information

- a Data Object can be either a Physical Object or a Digital Object. An example of the former is a piece of paper or a rock sample
- a Digital Object is made up of one or more Bits
- a Data Object is interpreted using Representation Information
- Representation Information is itself interpreted using further Representation Information because it is itself an Information Object which will have a Data Object and its own Representation Information.

Archival Information Package

For long term preservation an Archival Information Package (AIP) must be (logically) created, containing all the elements needed for preservation (here we use the AIP from [2] which includes Access Rights as part of PDI). This is shown in Figure 2.

As will be seen, besides Representation Information there is also Preservation Description Information (PDI), Packaging Information and Package Description.

Moreover CASPAR has developed [3] a set of workflows which complement the static view of the AIP.

Preservation Information Flow

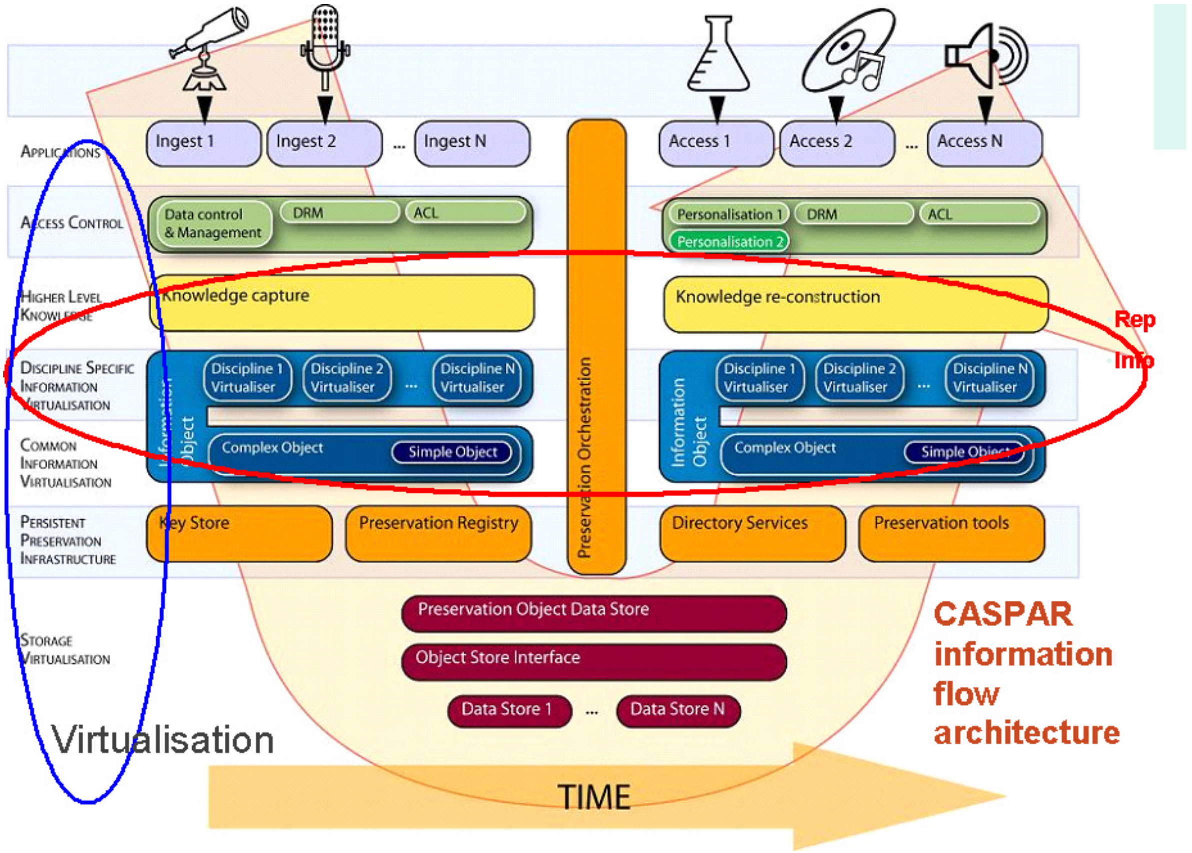

Figure 3 CASPAR Information Flow

Figure 3 shows one view that CASPAR has of the flow of information over time.

Many details must be captured as a Data Object comes into an archive, including

- access rights, Digital Rights Management (DRM) and Access Control Lists (ACL)
- various types of PDI (not shown in Figure 3
- Representation Information of various types
  - high level knowledge
  - various types of descriptions including a the way in which complex objects may be viewed as a composite of simpler objects. Some of these objects may be discipline specific whereas others are rather general.

For example an image is a fairly general concept – essentially an array of numbers, whereas an Astronomical image is an image plus an astronomical co-ordinate system and a way to map to physical measurements.

Details of the simple objects down to the bit level must also be captured.

Note that here, as well as elsewhere, virtualisation techniques can be applied. Further details of this and many other aspects of preservation can be found on the CASPAR web site and in particular the CASPAR Conceptual Model [3].

- The digital objects must be stored, indicated here as a Preservation Object Data Store.

Subsequently the process must be reversed when the Data Object (possibly after various Transformations, is needed for use and is taken out of storage, for example:

- Information must be extracted using the Representation Information at various levels
- Access constraints must be understood and respected

It is worth noting that much of these descriptions and extra pieces of information (metadata) will themselves be digitally encoded and will therefore also need to be preserved, using the same techniques.

What can change?

We can consider some of the things can change over time and hence against which an archive must safeguard the digitally encoded information.

- **Hardware and Software Changes**

  Use of many digital objects relies on specific software and hardware, for example applications which run on specific versions of Microsoft Windows which in turn runs on Intel processors. Experience shows that while it may be possible to keep hardware and software available for some time after it has become obsolete, it is not a practical proposition into the indefinite future, however there are several projects and proposals which aim to emulate hardware systems and hence run software systems.

- **Environment Changes**

  These include changes to licences or copyright and changes to organisations, affecting the usability of digital objects. External information, ranging from the DNS to DTDs and Schema, vital to the use and understandability, may also become unavailable.

- **Termination of the Archive**

  Without permanent funding, any archive will, at some time, end. It is therefore possible for the bits to be lost, and much else besides, including the knowledge of the curators of the information encoded in those bits. Experience shows that much essential knowledge, such as the linkage between holdings, operation of specialised hardware and software and links of data files to events recorded in system logs, is held by such curators but not encoded for exchange or preservation. Bearing these things in mind it is clear that any repository must be prepared to hand over its holding – together with all these pieces tacit of information – to its successor(s).

- **Changes in what people know**

  As described earlier the Knowledge Base of the Designated Community determines the amount of Representation Information which must be available. This Knowledge Base changes over time.

Preservation Strategies

It is sometimes argued [4] and [8] that the two preservation strategies available are emulation and migration. In fact there are a number of strategies which may be adopted, in particular the creation of additional Representation Information. Details are available on The CASPAR site and in particular in the CASPAR Validation report (http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-validation-evaluation-report/at_download/file).

Representation Information includes the description of the structure and the semantics of the digitally encoded object. CASPAR is developing and bringing together many techniques for producing and validating this type of description. Amongst the tools are ones for creating formal descriptions of Structure Information, for example as EAST or DRB descriptions. Semantic Information Data Dictionaries using DEDSL or ontologies, for example in CIDOC, are also needed. Further details are available in [3].

The question of how much Representation Information and whether it is adequate is addressed in OAIS through the concept of Designated Community. Ways to formalize this are discussed below. CASPAR has demonstrated techniques for validating the types and quantity of Representation Information by parsing the data using the descriptions, analogous to the way in which XML is validated. In addition we use the descriptions in generic applications to show to the satisfaction of the data experts that one can process and analyse the data object and produce the same results as with the software normally associated with it . Note that these generic applications are not meant as to replacement data specific current applications not least because the generic applications are slower and have limited functionality.

TOOLS AND INFRASTRUCTURE

All of these need to be created and maintained. CASPAR provides a number of toolkits for creating this information and a set of Key Components for maintaining them.

Figure 4 contains a number of workflows of importance for preservation. The key components of infrastructure are those components which are essentially independent of the information being preserved and therefore can be used for all types of information. The toolkits tend to be more data type dependent.

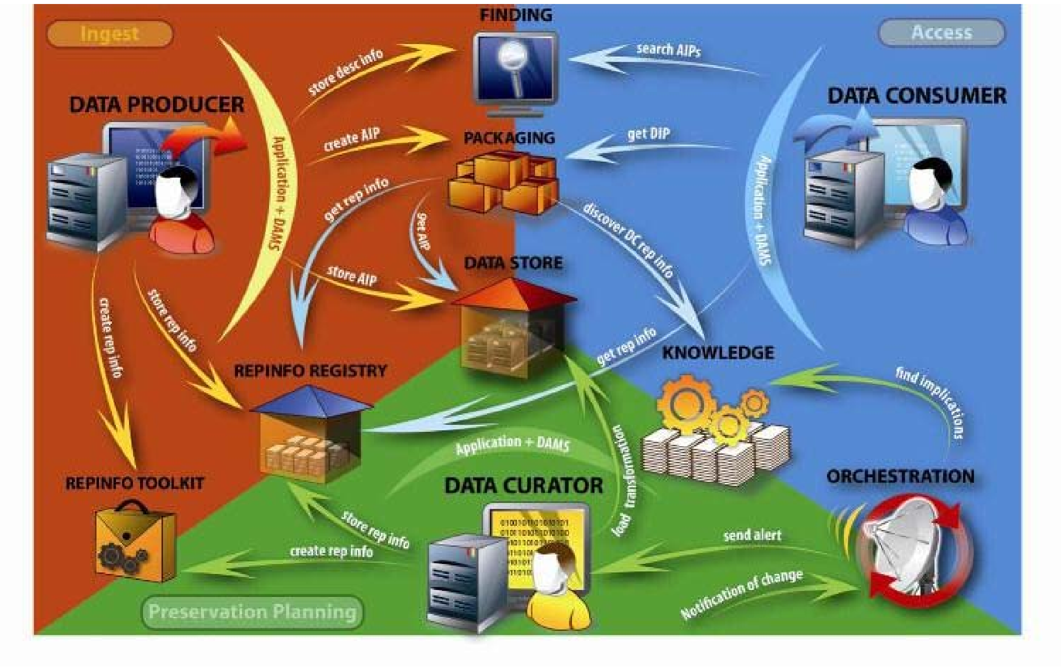

Figure 4 CASPAR Preservation Workflows

Mapping solutions to threats

Threat	General description of solution	CASPAR contribution
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	**RepInfo toolkit**, **Packager** and **Registry** – to create and store Representation Information. In addition the **Orchestration Manager** and **Knowledge Gap Manager** help to ensure that the RepInfo is adequate.
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	**Registry** and **Orchestration Manager** to exchange information about the obsolescence of hardware and software, amongst other changes. The Representation Information will include such things as software source code and emulators.
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	**Authenticity toolkit** will allow one to capture evidence from many sources which may be used to judge Authenticity.
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	**Digital Rights and Access Rights tools** allow one to virtualise and preserve the DRM and Access Rights information which exist at the time the Content Information is submitted for preservation.

Loss of ability to identify the location of data	An ID resolver which is really persistent	**Persistent Identifier** system: such a system will allow objects to be located over time.
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	**Orchestration Manager** will, amongst other things, allow the exchange of information about datasets which need to be passed from one curator to another.
The ones we trust to look after the digital holdings may let us down	Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term	The **Audit and Certification** standard to which CASPAR has contributed will allow a certification process to be set up.

One can then position these components and toolkits in a layered architecture, building on top of the network and e-Science infrastructure.

There are many levels and types of science data infrastructures, ranging from those specific to a domain of knowledge or organisation to, at the other extreme, something which would be generic, applicable to many domains and organisations. An analogy may be drawn between networks within an organisation and the network infrastructure which connects these into a network of networks. Similarly the distributed computing infrastructure bridges the gaps between islands of resource which contain compute and storage resources.

We know there is a complex system of repositories and this allows us to think about an "ecosystem" of repositories, in other words their size, distribution, interactions, and perhaps also their evolution.

We also know that individual repositories come into existence and change relatively rapidly to respond to changing user demands. Similarly the infrastructures within organisations and between repositories show invention and dynamism. All these are fundamental to facilitate the advancement of science.

We must in addition think of a broader science data infrastructure which is something which links together islands of resources and capabilities, some of which are themselves managed collections of resources. The principle of subsidiarity should apply when defining this broader infrastructure. The interfaces to such an infrastructure would change less rapidly in order to provide some stability in the system. Moreover the main function of such a generic infrastructure (linking together, for example, individual discipline infrastructures) would be to enable multi-disciplinary studies so that any scientist would be able to access and use/understand data from any other discipline, obtained at any time. Note that such a scientist would not expect this to require zero intellectual effort, but he/she should have minimal difficulties in, for example, getting hold of the digitally encoded information (we use the term data to include all types of digitally encoded information from static instrument data to dynamic databases to digital documents below) and the associated metadata.

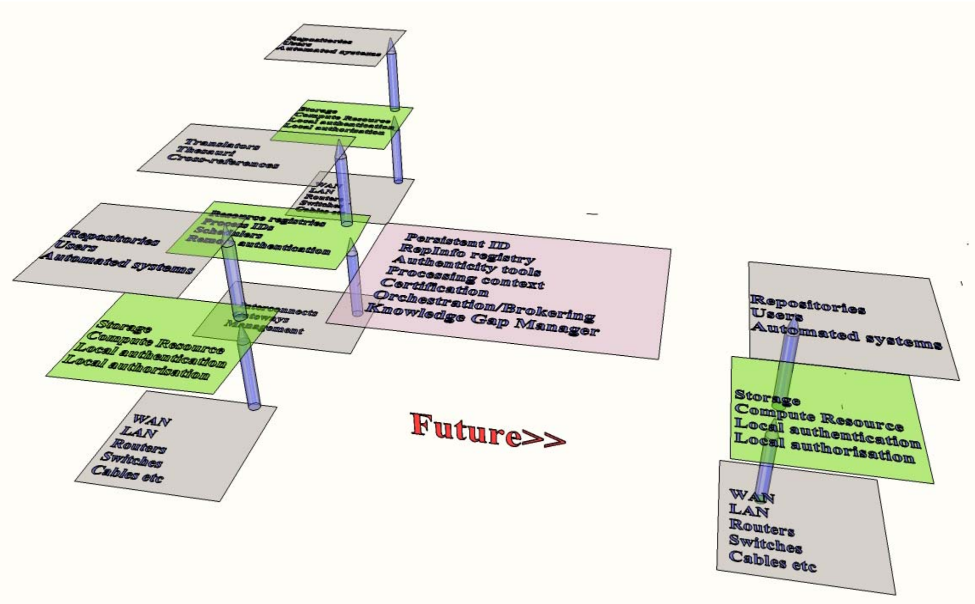

Figure 5 Infrastructure components

CONCLUSION

In order to maintain the understandability of digitally encoded information there is a need to provide mechanisms to allow people and organisations to share the burden over time. The PARSE.Insight project has identified a number of threats to digital preservation which must be countered. The work undertaken by CASPAR attempts to provide components and tools which provide solutions to these threats.

REFERENCES

[1] CCSDS. (2002). Reference model for an Open Archival Information System (OAIS). Retrieved on June 14, 2007 from the Consultative Committee for Space Data Systems (CCSDS) website: http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] OAIS update (at the time of writing under CCSDS review), http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[3] CASPAR D1201: Conceptual Model – Phase 1, (2007), Retrieved from http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-guidelines/at_download/file

[4] Granger,S (2000) Emulation as a Digital Preservation Strategy, D-Lib Magazine, October 2000. Retrieved from http://www.dlib.org/dlib/october00/granger/10granger.html

[5] Tzitzikas, Y. (2007). Dependency management for the preservation of digital information. 18th International Conference on Database and Expert Systems Applications, DEXA'2007. Regensburg, Germany, September 2007.

[6] Tzitzikas, Y., & Flouris, G. (2007). Mind the (Intelligibility) Gap. 11th European Conference on Research and Advanced Technology for Digital Libraries, ECDL'2007. Budapest, Hungary, September 2007.

[7] CASPAR Access Model, http://www.casparpreserves.eu/Members/cclrc/Deliverables/report-on-oais-access-model/at_download/file especially section 2.

[8] Stephan Strodl, Christoph Becker, Robert Neumayer, Andreas Rauber (2007), How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries table of contents, Vancouver, BC, Canada http://www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

[9] David Giaretta, Brian Matthews, Juan Bicarregui, Simon Lambert, Mariella Guercio, Giovanni Michetti and Donald Sawyer (2009), "*Significant Properties, Authenticity, Provenance, Representation Information and OAIS*" at iPRES 2009.

ACKNOWLEDGEMENTS

This paper reports work carried out by the many members of the PARSE.Insight and CASPAR projects, all of whom must be thanked for their efforts.