Sharing the Preservation Burden – 8375

David Giaretta

STFC, Rutherford Appleton Laboratory, Didcot, Oxon OX11 0QX, UK[1]

ABSTRACT

Preserving digitally encoded information which is not just to be rendered, as a document, but which must processed, like data, is even harder than one might think, because understandability of the information which is encoded in the digital object(s) is what is required. Information about Nuclear Waste will include both documents as well as data. Moreover one must be able to understand the relationship between the many individual pieces of information. Furthermore the volume of information involved will require us to allow automated processing of such information.

Preserving the ability to understand and process digitally encoded information over long periods of time is especially hard when so many things will change, including hardware, software, environment and the tacit and implicit knowledge that people have. Since we cannot predict these changes this cannot be just a one-off action; continued effort is required. However it seems reasonable to say that no organization, project or person can ever say for certain that their ability to provide this effort is going to last forever. What can be done? Can anything be guaranteed? Probably not guaranteed – but at least one can try to reduce the risk of losing the information.

We argue that if no single organization, project or person can guarantee funding or effort (or even interest), then somehow we must share the "preservation load", and this is more than a simple chain of preservation consisting of handing on the collection of bits from one holder to the next. Clearly the bits must be passed on (but may be transformed along the way), however something more is required – because of the need to maintain understandability, not just access. This paper describes the tools, techniques and infrastructure components which the CASPAR project is producing to help in sharing the preservation burden.

INTRODUCTION

CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) is an EU Integrated Project, which began in April 2006 (see http://www.casparpreserves.eu). It has total funding of about 16m Euro (8.8m Euros from the EU), and aims squarely at the problem of how to preserve digitally encoded information, and in particular what infrastructure will allow us to share the effort. CASPAR categorises the mechanisms for degradation of the understandability of information, including the disappearance of required hardware and software, unavailability of things from the general environment such as resources currently available on the internet, and

[1] This paper represents joint work of the CASPAR partners and special thanks must be given to: Luigi Fusco (ESA/ESRIN), Seamus Ross, (HATII, U of Glasgow), Mariella Guercio (U of Urbino), Mario Hernandez (UNESCO), Ugo Di Giammatteo (ACS), Zavisa Bjelogrlic (ASemantics), Dalit Naor (IBM/Haifa), Carlo Meghini (CNR), Silvia Boi (MetaWare), Daniel Terrgui (INA), Kia Ng (U of Leeds), Luigi Briguglio (Engineering), Vassilis Christophides (FORTH), Bruno Bachimont (CNRS/UTC), Hugues Vinet (IRCAM) and Pavel Sedlak (CIANT)

indeed those things which are currently "common knowledge" but which drop out of that category, for example changes in meaning of terminology.

This paper describes the tools, techniques and infrastructure components which CASPAR is producing to help in sharing the preservation burden. It will also touch on the ways in which the repositories which are the (temporary) custodians of the digitally encoded information can be tested or certified as trustworthy for long term custodianship. It will also discuss how this infrastructure can itself be embedded and be made preservable (i.e. accessible and usable) over the long term.

SOLUTIONS OR SNAKE OIL?

It is easy to propose some solutions – and extremely easy to wave one's hands. The difficulty is to provide evidence of effectiveness - other than simply waiting a long time! This in a sense brings us to the CASPAR acronym in that the reason the project includes science, arts and culture and a number of other disciplines is that there is a need to test what is done, and test it "for real" in a variety of scenarios involving science data from ESA and CCLRC, Cultural Heritage data from UNESCO and Performing Arts data from IRCAM, Univ. Leeds, INA and CIANT.

It is, for example, relatively easy to claim that the solution is to write everything out as XML – but how can that be verified? One may claim that a technique, for example emulation, works as can be shown for a certain example, but does it work for all types of digitally encoded information? What does the claim "I am preserving this digital object" mean?

OAIS Reference Model

The OAIS Reference Model (ISO 14721)(1) is one of the most important standards in this area, providing a number of important concepts and terminology. Its view of digital preservation is very general, but in fact its approach means that digital preservation is even harder than one might think, because it talks in terms of understandability of the information which is encoded in the digital object(s).

It could be argued that one could, for example, make a "digital" object by carving 1's and 0's in stone – a very durable way to preserve information as the ancient Egyptians knew. However while this may give one access (slow access - but access nevertheless) – it will not maintain understandability, as shown by the example of the Phaistos disk (dated to 1700 BC) which has still not been translated.

For example, in a particular file, even if one can extract a number from that file, what does it mean; what is the relationship to other numbers in that file or in other files? In order to understand digitally encoded information a swarm of additional information (various types of metadata, and in particular Representation Information in OAIS terminology) is required.

The OAIS approach is essentially that there must be a way of testing any claims of preservation and the criterion is that the information must remain understandable and usable. This then brings in the concept of Representation Information, defined as *information that maps a Data Object into more meaningful concepts*, shown in the following diagram.

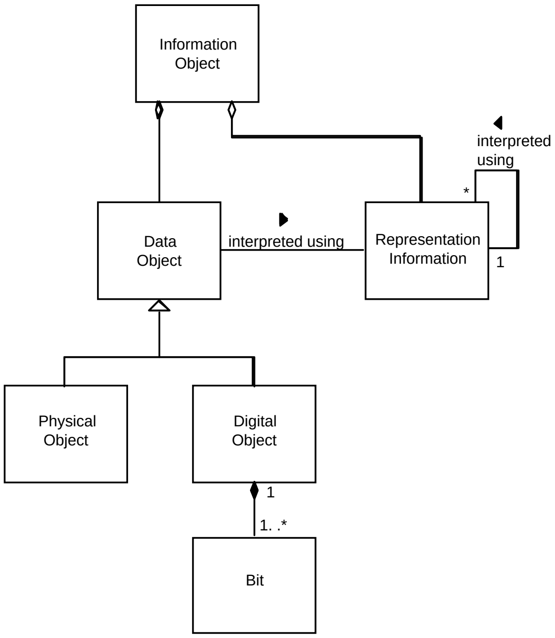

Figure 1 OAIS Information Model

For those unfamiliar with UML notation the following key, taken from the OAIS Reference Model document (1) should be helpful.

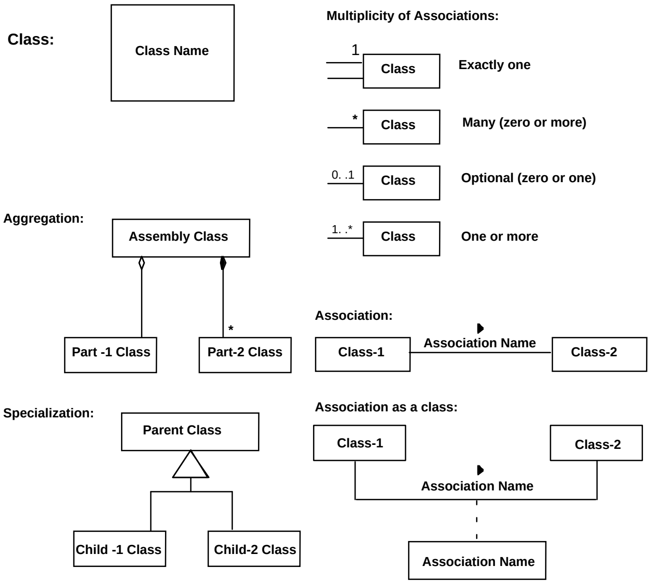

Figure 2 Key to UML Relationships

Representation Information is a catch-all concept covering essentially everything that is needed to make a particular collection of bits (the Content Data Object) understandable and usable. However simply saying that one needs Representation Information is not enough; OAIS recognises that Representation Information itself is captured as a Data Object which itself needs its own Representation Information. The Representation Network which OAIS defines as the *set of Representation Information that fully describes the meaning of a Data Object*, is a very large collection, as the example in the box below indicates. The purpose of the box is to illustrate the types of Representation Information which might be ultimately be needed when the people for whom the data is being preserved do not share a good deal of what is currently common knowledge.

Representation Information for Martians

How much Representation Information would one need to provide for a Martian to understand and use Ionosonde (see http://en.wikipedia.org/wiki/Ionosonde) data which is digitally encoded?

Where to start? Let’s start with a definition on paper of the format – and maybe a Rosetta Stone equivalent of Martian to English (or Chinese or whatever language the document is written in). But what about some other things like bits? binary notation, IEEE encoding for floating point numbers, definitions of the names of the data values, relationship between the data values, definition of frequency, definition of a second, basic physics, graduate-level physics, English, etc etc? The list is very, very long.

In order to provide a way of limiting the size of the Representation Network, OAIS introduces the concept of **Designated Community**. This is *an identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.* However this only holds back the flood of Representation Information temporarily.

The difficulty is that what the Designated Community knows (their Knowledge Base), and therefore does not have to be provided as Representation Information, changes over time. In other words even if one could engrave the binary sequences on stone, or something else as permanent, that is only part of the job – that only guarantees continued access, but **not** continued understandability.

It is worth emphasising this distinction further by consider the case where one has access to a bit sequence which encodes information about concentrations of nuclear material in number of containers. At one extreme the data may be encrypted so that having the bits is not very useful unless one can crack the encryption. At another extreme the bit sequence may be recognised as an ASCII encoded comma separated value (CSV) file of the type which may be imported into spreadsheet applications. In this second case one may be able to see the numbers and character strings in the file, but if one does not have, for example, the measurement units associated with the numbers one will have to guess whether one is dealing with a very large or very small risk

VALIDATION METRICS

CASPAR proposes a number of rather general metrics for validating itself and these metrics should, with minor changes, be applicable to most other claims about digital preservation techniques (2). These may be summarised as:

- demonstrate a sound theoretical basis for the approach taken
- practical demonstration by means of what may be regarded as “accelerated lifetime” tests involving:
    - software, hardware and environment changes

  - changes in the Designated Communities and their Knowledge Bases
- show improved trustworthiness of repositories – for example using the work on audit and certification coming from the Repository Audit and certification group (3) which aims to produce an ISO standard on which accreditation and certification processes for digital repositories can be based.

It is fair to say that these cannot provide **absolute** proof – only **evidence** to support the claim of effectiveness, in that there are risks involved in long term preservation about issues ranging from availability of electrical power to continued social stability, which, even though small, mean that absolute guarantees cannot be given.

SHARING THE BURDEN

Returning to the issue of how to share the burden of preservation, an analogy may be drawn to the Wikipedia which has many, many contributors producing/correcting/moderating content – and which has become one of the most authoritative (or at least most Googled) sources of information on the Internet. Similar efforts of harnessing multiple contributors are going on in the BBC, which is setting up something which has been described as "wiki-radio" (4) where the public can annotate recorded material. Another example is Google which also relies not on its own judgement on the value of a page but rather on the value other place on that page – the page ranking algorithm. Books such as "The wisdom of crowds" describe many more such examples of harnessing that "wisdom".

However digital preservation needs to more than just the equivalent of a Wiki; there is a need to be proactive, in other words not simply relying on individuals to contribute ideas but instead actively to prompt people for this input, otherwise the information will be lost through neglect, and CASPAR's preservation infrastructure components are intended for just this purpose. The underlying architecture is available in much greater detail in the CASPAR Conceptual Model (5). In brief the infrastructure must include components to allow us to:

- collect the contributed "wisdom" – the Registry/Repositories
- remind people to take action – the Representation Information Gap Manager and the Orchestration Manager
- capture the "wisdom" using "local" tools for creation of Representation Information, Preservation Description Information etc, as well as techniques for Persistent Storage

PRESERVATION INFRASTRUCTURE

Key components in a preservation infrastructure need in particular to facilitate the capture and use of Representation Information. While any repository of digital information about, for example, Nuclear Waste, will wish to keep its information, and the associated Representation Information, under close control, nevertheless there needs to be a mechanism for bringing in the "wisdom." The techniques which CASPAR proposes are described in terms of a distributed system, but one would imagine periodic incorporation of contributed, initially distributed, information into a central repository.

The underlying idea is illustrated in Figure 3..

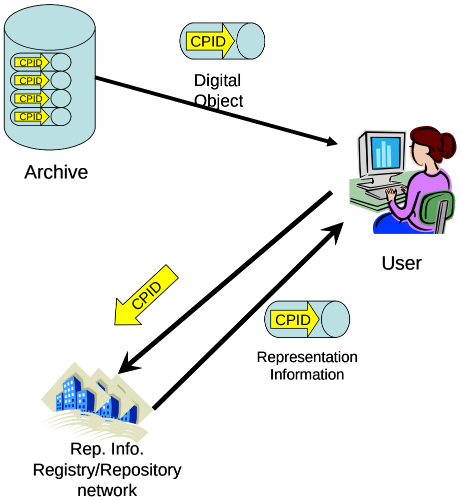
CPID
Archive
CPID
Digital Object
User
CPID
CPID
Representation Information
Rep. Info. Registry/Repository network

Figure 3 Use of Registry/Repository of Representation Information

In this figure the steps are:

- a user requests a piece of digitally encoded information from an archive
- the archive provides the data object, perhaps including some Representation Information (RepInfo), plus a pointer (CPID) to additional RepInfo
- if the user finds that there is insufficient Representation Information packaged with the data from the repository then the CPID is used to access additional RepInfo from a Registry/Repository
- the Registry/Repository returns the RepInfo requested, which in turn has an associated CPID which points to its own RepInfo.

The above is not meant to imply that there must be a single, unique, Registry/Repository, nor even a single definitive piece of Representation Information for any particular piece of digitally encoded information.

Filling in the gaps that arise

Gaps will arise over time between the level to which there is available explicit Representation Information and the level required by users, as software, hardware, environment and the knowledge base of the designated community changes. In order to identify what additional Representation Information must be captured in order to fill these gaps the Registry/Repository is supplemented by the Knowledge Manager – more specifically a Representation Information Gap manager which identifies these gaps. Of course the information on which this is based does not come out of thin air. People (initially) must provide this information and an Orchestration Manager collects and distributes this information.

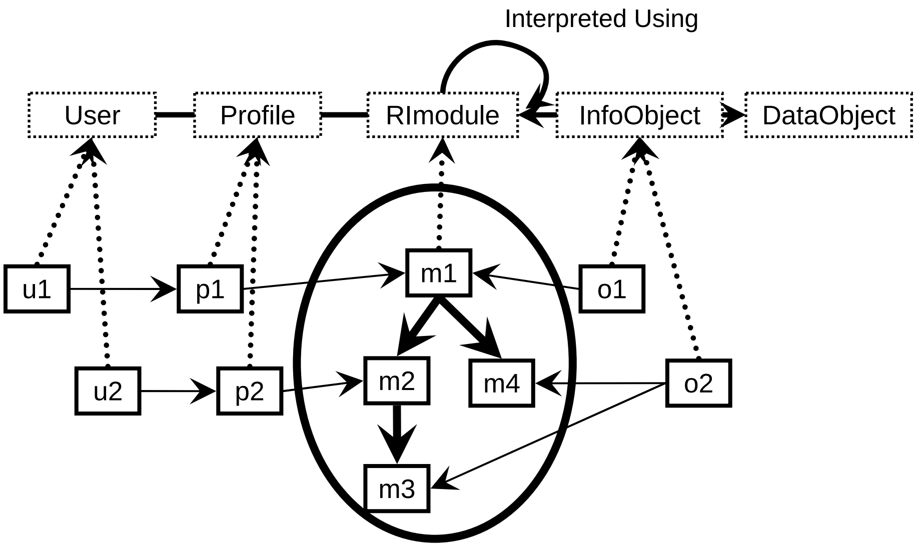

Figure 4 Modelling Users, Profiles, Modules and Dependencies

Support for automation in identifying such “gaps” , based on information received, is illustrated in Figure 4 which shows users (***u1, u2***…) with user profiles (***p1, p2***… – each a description of the user’s Knowledge Base) with Representation Information {***m1, m2***,…) to understand various digital objects (***o1, o2***…).

Take for example user ***u1*** trying to understand digital object ***o1***. To understand ***o1***, Representation Information ***m1*** is needed. The profile ***p1*** shows that user ***u1*** understands ***m1*** (and therefore its dependencies ***m2, m3*** and ***m4***) and therefore has enough Representation Information to understand ***o1***.

When user ***u2*** tries to understand ***o2*** we see that ***o2*** needs Representation Information ***m3*** and ***m4***. Profile ***p2*** shows that ***u2*** understands ***m2*** (and therefore ***m3***), however there is a gap, namely ***m4*** which is required for ***u2*** to understand ***o2***.

For ***u2*** to understand ***o1***, we can see that Representation Information ***m1*** and ***m4*** need to be supplied.

This illustrates one of the areas in which Knowledge Management techniques are being applied within CASPAR, in addition to the capture of Semantic Representation Information.

A formal treatment of these ideas are available (6, 7).

AUTOMATION AND BANG FOR THE BUCK

A perfectly acceptable form of Representation Information could be simply a (probably huge) paper document describing all aspects of how to get information out of the bit sequences. If such a document is the only Representation Information available then it is clearly better than having no Representation Information.

However one important drawback of this type of Representation Information is that it is difficult to use; it requires (at the moment) a human to read and understand it. This is almost certainly relatively slow and expensive, and difficult for someone to do when there are hundreds or thousands of different types of data objects to deal with.

CASPAR aims, where possible, to create types of Representation Information which supports automation, in other words which are likely to be usable in tools and software which are available in the future.

The Warwick Workshop (8) noted that Virtualisation is an underlying theme; however, virtualisation is not a magic bullet. It cannot be expected to be applied everywhere, and even where it can be applied the interfaces can themselves become obsolete and will eventually have to be re-engineered/re-virtualised, nevertheless we believe that it is a valuable concept. Each of these levels of virtualisation will have its own type of “virtualisation description”, which is a type of Representation Information, which will also need its own Representation Information.

Digital preservation is about using what will by then be unfamiliar digital objects in the future. In many ways e-Science (or GRID) is about using digital objects right now, irrespective of whether those objects were created centuries or seconds ago, and these digital objects are likely to be unfamiliar simply given the number of sources of information which are becoming available. In CASPAR we argue that the Representation Information which supports automation and which is gathered for preservation also supplies a need in e-Science, namely that of making collections of bits into information which can be dealt with in an automated way.

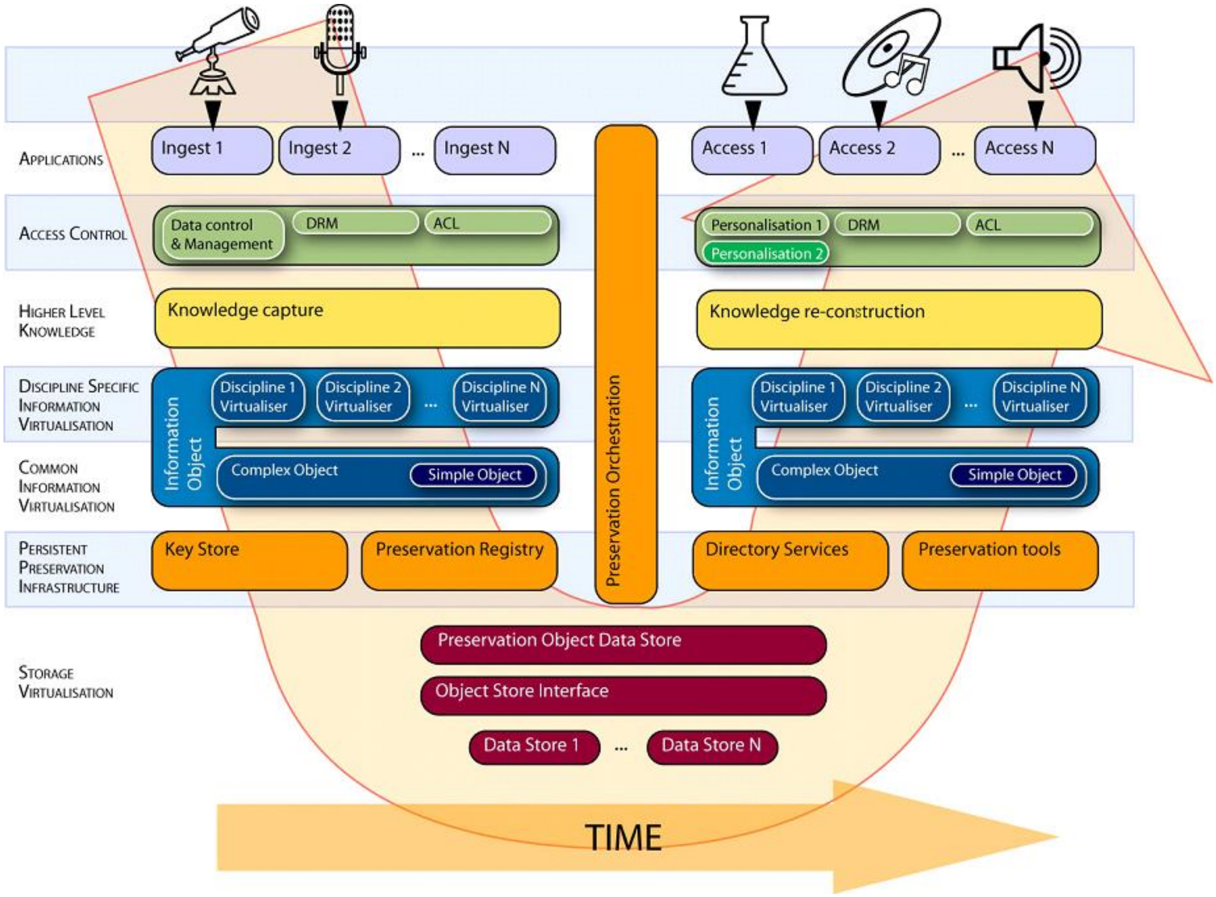

Figure 5 CASPAR Information Flow Architecture

Figure 5 indicates in somewhat more detail than a number of layers in which CASPAR expects to use Virtualisation including:

- Digital Object Storage virtualisation - which extends ideas related to distributed storage so that the Digital Object storage handles whole objects rather than byte segments, and further adds curation capabilities such as automated tracking of certain types of provenance.
- Common information virtualisation – which will facilitate use of data in new (as yet unknown) applications by explicitly showing how digital objects (the bit sequences) can be treated as one or more of the archetypal information objects such as image, table, tree or document.
- Discipline specific information virtualisation – for example extending the simple objects such as "image" to specialisations such as multi-spectral image, Earth Observation image, nuclear waste maps, where what one adds are specialised functionality which are agreed that all waste maps, for example, should be able to support.

Virtualisation also applies to the

- Higher level knowledge
- Access control and Digital Rights Management
- Processes

Each of these layers must be preserved, and this requires that each of the artefacts, which are created at each layer, will itself be a digital object, and must itself be preserved.

SUMMARY

CASPAR is attempting to use OAIS concepts rigorously and to the fullest extent possible, supplementing these where appropriate. Based on these fundamental ideas about digital preservation, a number of components, tools and techniques are being created in order to provide a broadly applicable infrastructure to allow the spreading of the burden of preserving the understandability and usability of digitally encoded information.

In the process the limits of the applicability of these OAIS concepts are themselves being tested. Most importantly a number of validation metrics have been produced. Further details are available from the CASPAR website http://www.casparpreserves.eu.

REFERENCES

1. OAIS Reference Model http://public.ccsds.org/publications/archive/650x0b1.pdf
2. CASPAR Description of Work http://www.casparpreserves.eu/Members/metaware/ReferenceDocuments/caspar-description-of-work/at_download/file Table 1 - Digital Preservation Metrics.
3. http://wiki.digitalrepositoryauditandcertification.org [viewed 1 Dec 2007]
4. See for example http://www.bbc.co.uk/radio4/science/findlistenlabel/?programme=allinthemind20070410 [viewed 11 June 2007]

5. http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-conceptual-model-phase-1-1/at_download/file [viewed 11 June 2007]

6. Y. TZITZIKAS, “Dependency Management for the Preservation of Digital Information", 18th International Conference on Database and Expert Systems Applications, DEXA’2007, Regensburg, Germany, September 2007

7. Y. TZITZIKAS and G. FLOURIS, “Mind the (Intelligibility) Gap", 11th European Conference on Research and Advanced Technology for Digital Libraries, ECDL’2007, Budapest, Hungary, September 2007

8. http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf