iPRES 2009

THE SIXTH INTERNATIONAL CONFERENCE ON THE PRESERVATION OF DIGITAL OBJECTS

Proceedings

October 5-6, 2009
Mission Bay Conference Center
San Francisco, California

CDL

California Digital Library

Tools for Preservation and Use of Complex and Diverse Digital Resources

David Giaretta
Science & Technology Facilities Council, Rutherford Appleton Lab., Didcot, Oxon, UK and Director of the CASPAR project

Abstract

To preserve digitally encoded information over the long term following the OAIS Reference Model requires that this information remains accessible, understandable and usable by a specified Designated Community. These are significant challenges for repositories, particularly when dealing with scientific data where the semantics must be associated with the data so that one can answer the basic questions such as "what does this number actually measure?". Answering these basic questions lead to techniques which fit naturally into a strategy for re-use of information across disciplines.

The CASPAR project (http://www.casparpreserves.eu) is an EU part funded project with total spend of 16MEuros which is trying to faithfully implement almost all aspects of the OAIS Reference Model in particular the Information Model. The latter involves tools for capturing all types of Representation Information (Structure, Semantics and all Other types), and tools for defining the Designated Community.

This paper will describe the tools and infrastructure components which have been implemented by the CASPAR project to support repositories in their task of long term preservation of digital resources. We address also the capture and preservation of digital rights management and evidence of authenticity associated with digital objects. Moreover examples of ways to evaluate a variety of preservation strategies will be discussed as will examples of integrating the use of these infrastructure components and tools into existing repository systems.

Examples will be given of a rich selection of digital objects which encode information from a variety of disciplines including science, cultural heritage and also contemporary performing arts. While there are many common aspects this selection also provides a rich set of distinct challenges both technically as well as organisationally. We will compare other preservation techniques which are in common use and show their deficiencies when dealing with these challenges. For example, even a relatively simple document may be a container of separate objects which have their own, say, provenance, and also have important and complex relationships to other objects both within and outside the containing document. We will demonstrate techniques for dealing with this and the many other challenges and compare them with other types of solutions from other projects.

Introduction

Preserving digitally encoded information over the long term is hard, as many studies and articles will confirm. The OAIS Reference Model [1] is one of the most important standards in this area and its view of digital preservation is very general, but in fact its approach makes it clear that digital preservation is even harder than one might think. To preserve a digital object requires effort which must be sustained over the long term.

It might 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, a point we will return to, is that while this may give one access (slow access but nevertheless it is access) – it will not maintain understandability.

Continued effort requires continued funding; it is reasonable to say that no organisation, project or person can ever say for certain that their funding is going to last forever (or even for the next 3 years). What can be done? Can anything be guaranteed? Probably not – but at least one can reduce the risk of losing the information.

Expanding on this rather obvious observation, we argue that if no single organisation, 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). This can involve duplicate copies – the analogy of multiple copies of books – and in the digital world we have for example LOCKSS. But something more is required – because of the need to maintain understandability, not just access.

This article describes some of the fundamental concepts in CASPAR as well as the metrics by which CASPAR believes that it, and other projects which claim to aid the practice of digital preservation, should be judged.

Further details are available from the CASPAR web site http://www.casparpreserves.eu; details of the software is available at http://developers.casparpreserves.eu:8080 and source code at http://www.sourceforge.net/projects/digitalpreserve

Two of the main aims of CASPAR are:

1. to produce tools and key infrastructure components to support digital preservation of digitally encoded information, strongly adhering to the concepts of the OAIS Reference Model [1] and its update [2].
2. to validate this infrastructure, in other words – can the project offer evidence that the tools and

techniques claimed to be effective for digital preservation are indeed effective.

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.

Representation Information

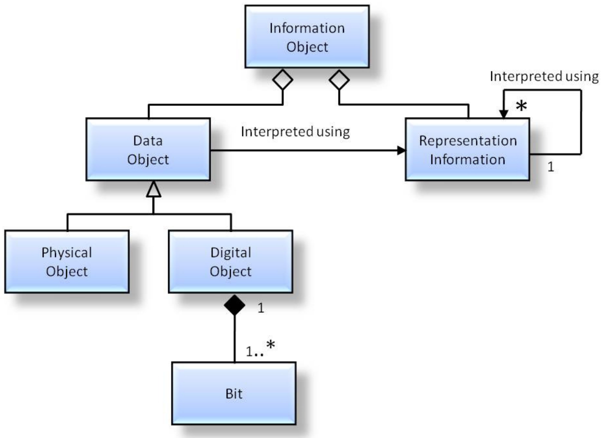

Figure 1 OAIS Information Model

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.

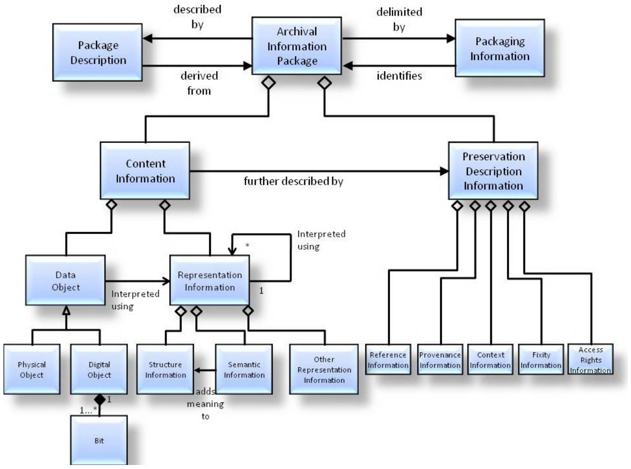

Figure 2 OAIS Archival Information Package (AIP)

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

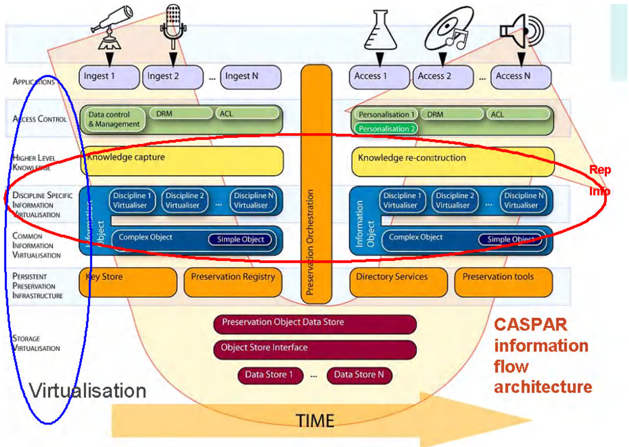

Figure 3 CASPAR Information Flow

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.

  - o 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.
  - o Details of the simple objects down to the bit level must also be captured.
  - o 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, each having its limitations as discussed next.

Preservation Strategy selection

Figure 4 shows the workflow appropriate for selecting preservation strategies which includes analyses of the archive, preservation objectives and, very importantly, identification of the Designated Community. Each strategy is evaluated in a cost/benefit analysis which is, it must be admitted, still quite rudimentary.

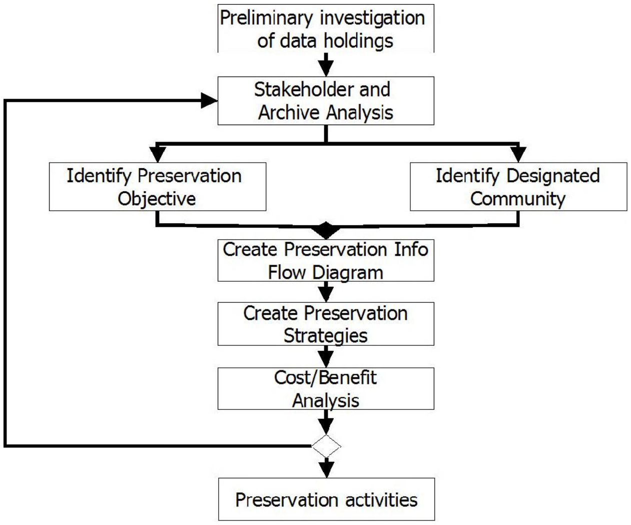
Preliminary investigation of data holdings
Stakeholder and Archive Analysis
Identify Preservation Objective
Identify Designated Community
Create Preservation Info Flow Diagram
Create Preservation Strategies
Cost/Benefit Analysis
Preservation activities

Figure 4 Preservation Strategy selection workflow

A number of these strategies are described next. The tools which are needed are described in section "Tools and Infrastructure".

Emulation

Emulation has been defined as "the ability of a computer program or electronic device to imitate another program or device." (Wikipedia, emulation, from http://en.wikipedia.org/wiki/Emulation. Retrieved July 25 2008). The fundamental aim is to do with current hardware and software what could be done in the past. This is adequate for rendering documents and running applications such as computer games on single machines. However it is very limiting in the context of data because at least sometimes (and perhaps most often) one does not simply reproduce what has been done previously; rather one wants to re-purpose and re-analyse archived data using modern tools and techniques. In addition emulation currently has difficulties with network aware applications and processes.

It is worth noting that emulation systems may be regarded as special types of Representation Information in that it may assist the understanding of digitally encoded information.

Access software

Access Software [1] "presents some or all of the information content of an Information Object in forms understandable to humans or systems. It may also provide some types of access service, such as displaying, manipulating, processing, or sub-setting, to an Information Object."

This allows one to "plug-in" to new software to access information encoded in digital objects. However there may be a mismatch in concepts. CASPAR addresses this through the application of virtualisation techniques which attempt to identify common and discipline related types of objects, for example images or tables, on the assumption that future systems will employ these concepts and so the "plug-in" is likely to be more easily and reliably developed.

Migration

The primary types of migration, ordered by increasing risk of information loss, are [1]:

- **Refreshment**: A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by copying the bits on the medium used to hold AIPs and to manage and access the medium. As a result, the existing Archival Storage mapping infrastructure, without alteration, is able to continue to locate and access the AIP.
- **Replication**: A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication may require changes to the Archival Storage mapping infrastructure.
- **Repackaging**: A Digital Migration where there is some change in the bits of the Packaging Information.
- **Transformation**: A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content. It is this type of migration which is usually referred to in the context of "emulate or migrate".

One important consideration is the need to analyse the implications of any type of migration, and in particular for repackaging or transformation. Regarding transformation one needs to think carefully, for example about:

- the potential loss of information – especially if special conventions have been adopted on top of particular formats, which are then embodied in access software. In addition the underlying information concepts which are captured in the initial and final forms may not match – here the virtualisation concepts from CASPAR can help.
- the associated costs of transformation, for example whether it should be done in bulk or on-demand.
- the implications for authenticity – how can one prove that the transformed version is indeed sufficiently the "same" as the original, as discussed in the accompanying paper [9]
- the preservability of the new form

Create Representation Information

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

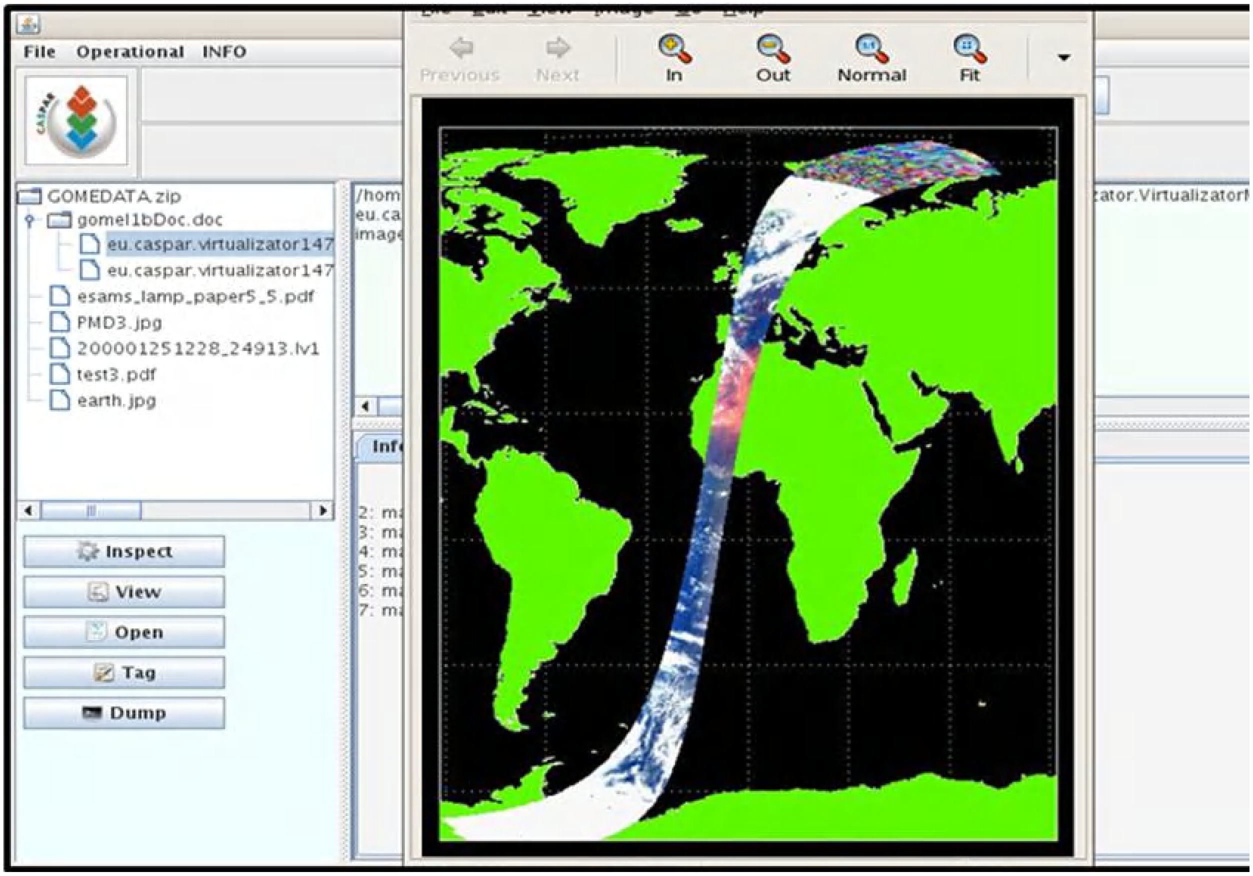

Figure 5 Delving into container files

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.

Toolkits

RepInfo Toolkit: a framework for many different tools for creating Representation Information. Besides tools for the creation of formal descriptions such as EAST, there are also tools for describing, and adding Semantic Information to, digital objects within containers. The simplest such container is a ZIP file. However a Word file can also contain, for example, images. Figure 5 shows an example of delving into a ZIP file which contains, amongst other things, a Word file. The Word file contains a table which contains images. One such image is shown and Semantic Information may be added, in an external description file.

Authenticity Toolkit: for capturing evidence which can be evaluated to judge Authenticity, and ensuring that it cannot be tampered with or denied. The accompanying paper on "Significant Properties, Authenticity, Provenance, Representation Information and OAIS" discusses some aspects of this. The Authenticity tool is based around the Authenticity Model which has been described elsewhere [7].

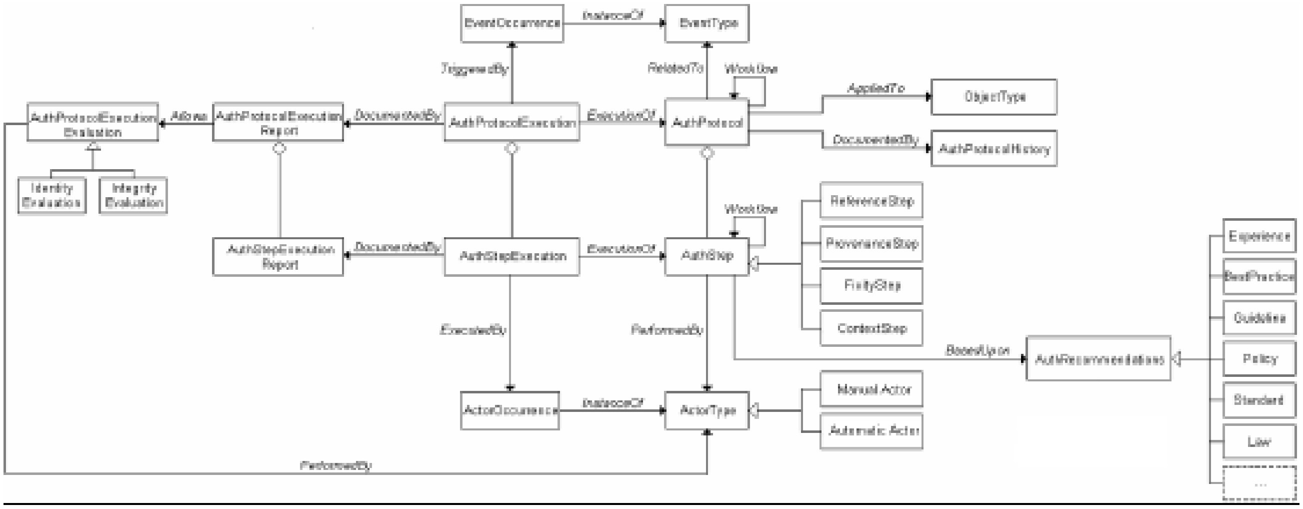

Figure 6 Authenticity Model

Access Rights Tools: for capturing Access Rights including Digital Rights is based on an Intellectual Rights ontology shown in Figure 8.

The Rights Ontology is harmonized with CIDOC CRM. As a consequence, the two Provenance data objects (the life cycle expressed in CIDOC and the rights expressed in terms of the Rights Ontology) are integrated (see Figure 7); this means that it is possible to navigate from one part to the other, and thus to implement search and retrieval of AIPs based on Provenance criteria, involving both rights and creation history.

It is important to capture properly and to preserve the creation history, and not just the rights themselves, because if something changes, like in the next scenario, then one must reconstruct the rights starting from the events that originated them.

The artifacts created by these tools are themselves for the most part digital objects which require preservation and therefore need their own Representation Information, Provenance, Fixity etc. For example the Digital Rights can take into account changes in the law and derive changes in rights associated with these objects.

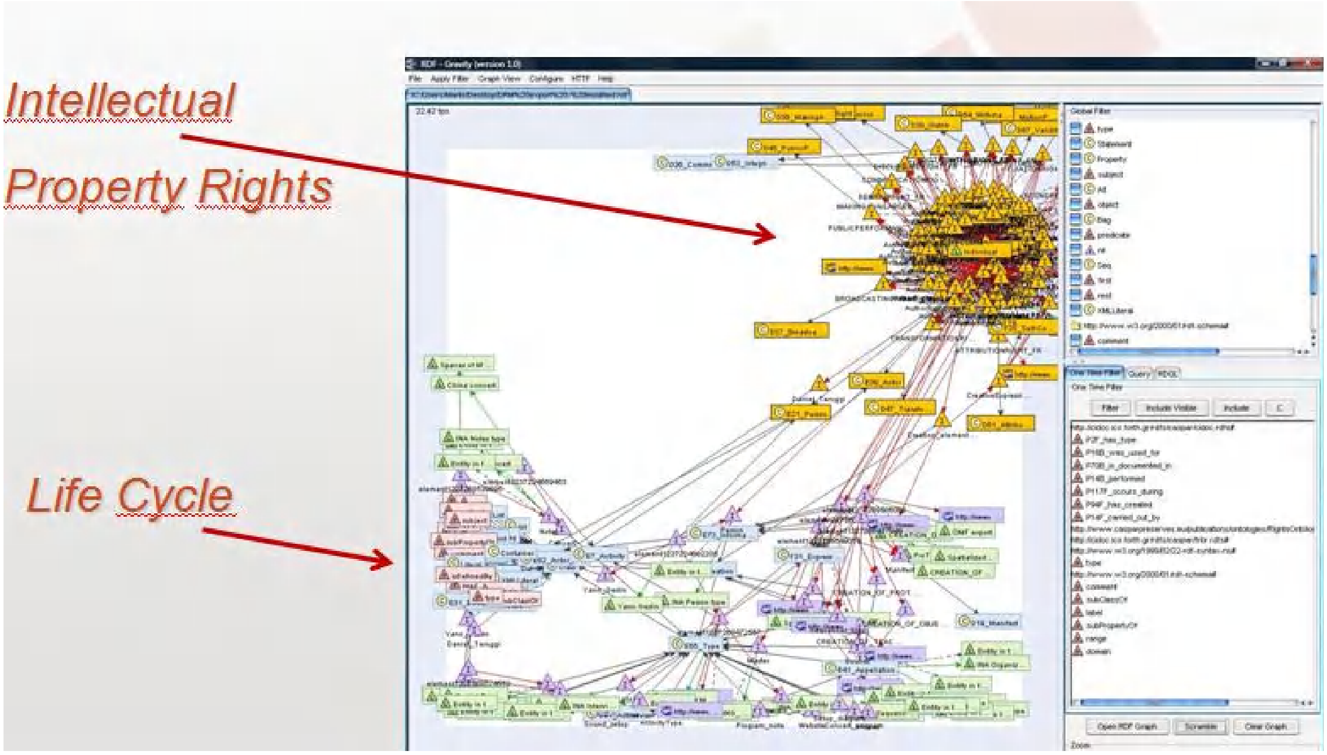

Figure 7 Intellectual Property and Provenance

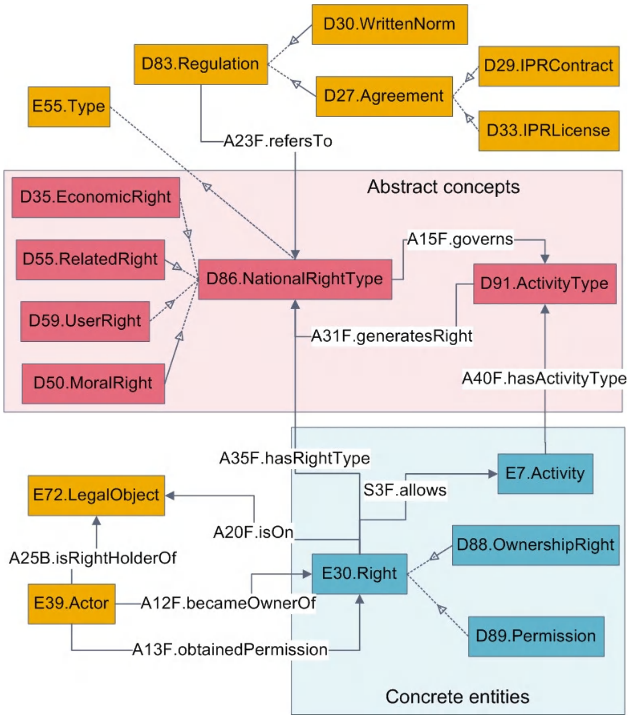

Figure 8 Digital Rights Ontology

The artifacts created by these tools are themselves for the most part digital objects which require preservation and therefore need their own Representation Information, Provenance, Fixity etc. For example the Digital Rights can take into account changes in the law and derive changes in rights associated with these objects.

Infrastructure components

Figure 9 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.

Two of the workflows are described in the next section.

Workflows for use of digital objects

The following workflow, extracted from Figure 9, illustrates the way in which digital objects may be used and understood by users.

The basic idea is that Representation Information must be associated with the Data Object. Identifiers (called here Curation Persistent Identifiers - CPID) which can be associated with any data object, point to the appropriate Representation Information in a Registry/ Repository, as illustrated in Figure 10. The Representation Information returned by the Registry/Repository itself is a digital object with its own CPID.

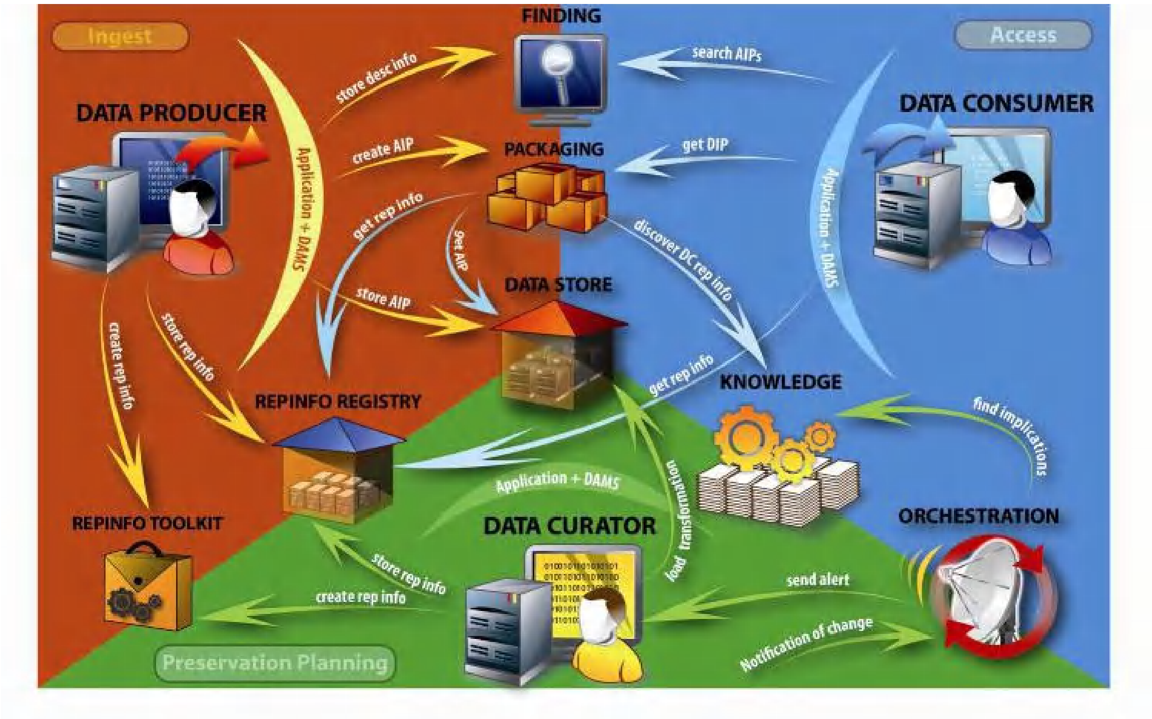

Figure 9 Preservation Workflows

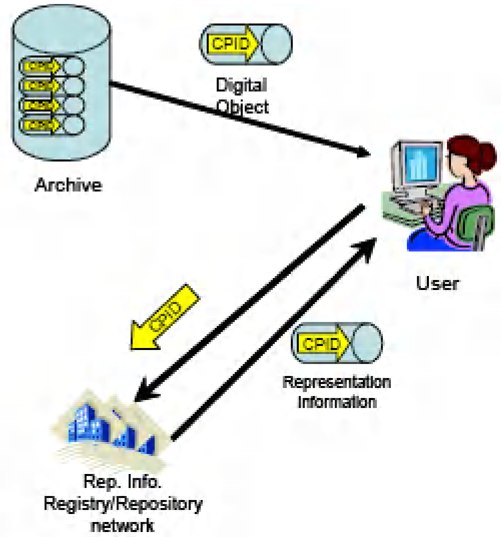
CPID
Digital Object
Archive
User
CPID
CPID
Representation Information
Rep. Info. Registry/Repository network

Figure 10 Use of Registry/Repository of Representation Information

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. The Representation Information may be packed with the Data Object or may be otherwise stored locally. The issue which must be considered next is maintaining the Representation Network. This is crucial because the allows the Data Object to remain understandable despite changes in hardware, software, environment and the Knowledge base of the Designated Community. As a result of these changes "gaps" will arise between the available Representation Network and the Designated Community's Knowledge Base. The way in which these are filled is addressed in the next section.

Workflows for Maintaining the Representation Information Network of Digital Objects

The Registry/Repository is supplemented by the Knowledge Manager – more specifically a Representation Information Gap manager which identifies gaps which need to be filled, based on information supplied to the Orchestration component.

Of course the information on which this is based does not come out of thin air. People (initially) must provide this information and the Orchestration Manager (Figure 11) collects this information and distributes.

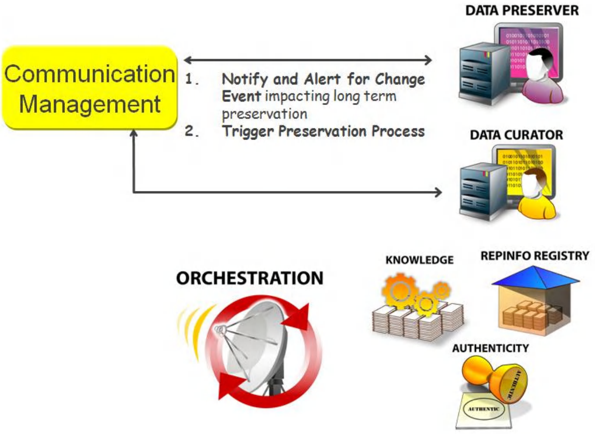
DATA PRESERVER
Communication Management
1. Notify and Alert for Change Event impacting long term preservation
2. Trigger Preservation Process
DATA CURATOR
ORCHESTRATION
KNOWLEDGE
REPINFO REGISTRY
AUTHENTICITY

Figure 11 Orchestration / Communication

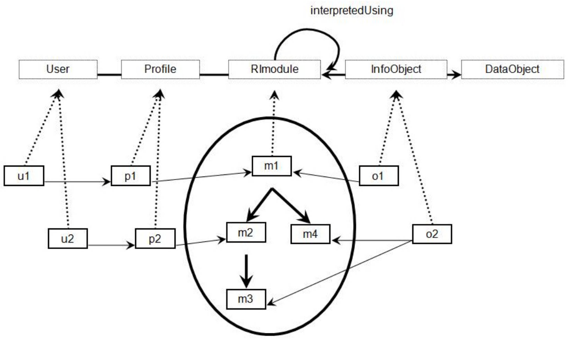

Figure 12 Representation Information dependencies

Support for automation in identifying such "gaps" , based on information received, is illustrated in Figure 12 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. Further details are available in the CASPAR Conceptual Model [3] and [5] and [6].

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

Validation Metrics

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 – the reason we have science, arts and culture (and more…) is that we need to test what we do, and test it "for real" in a variety of scenarios involving science data from ESA and STFC, Cultural Heritage data from UNESCO and Performing Arts data from IRCAM, University of 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?

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. These may be summarised as:

- demonstrate a sound theoretical basis for the approach taken
- provide practical demonstrations by means of what may be regarded as "accelerated lifetime" tests involving:
  1. hardware, software and environment changes
  2. changes in the Designated Communities and their Knowledge Bases
- show improved trustworthiness of repositories

It is fair to say that these cannot provide absolute proof of effectiveness – only evidence to support the claim of effectiveness.

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 work undertaken by CASPAR attempts to provide these critical components. There are many types of metadata which must be created and maintained and CASPAR has attempted to address and provide tools for all of these.

Digital preservation approaches which focus only on emulation and migration, or which address only formats, rather than including all types of Representation, in particular Semantic Representation Information, are able to address only a small part of the problem space and are usable only for limited aspects of preservation of a limited number of types of digitally encoded information.

Acknowledgements

This paper summarises part of the work of the CASPAR project and represents the efforts of many individuals to numerous to include in the list of authors. Full details are available on the CASPAR web site (http://www.casparpreserves.eu) .

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.