The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

The CASPAR[1] Approach to Digital Preservation

David Giaretta*,
Science and Technology Facilities Council (formerly CCLRC)
and the UK Digital Curation Centre

June 2007

Summary

A description of 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.

[1] Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project CASPAR contract IST-033572. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Preserving digitally encoded information over the long term is hard, as many studies and articles will confirm. The OAIS Reference Model is one of the most important standards in this area and its view of digital preservation is very general, but in fact its approach means that digital preservation is even harder than one might think.

Thinking rather generally about digital preservation, it is probably fair to say that to preserve a digital object requires effort which is sustained over the long term.

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, a point I 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 say for certain that existing funding is going to last forever (or even more than the next 3 or 7 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[2]. But something more is required – because of the need to maintain understandability, not just access.

CASPAR (**C**ultural, **A**rtistic and **S**cientific knowledge for **P**reservation, **A**ccess and **R**etrieval) is an EU Integrated Project, which began in April 2006. It has total funding of about 16m Euro (8.8m Euros from the EU), and aims squarely at this problem. 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.

The project brings together a consortium covering important digital holdings, with the appropriate extensive scientific (CCLRC (the lead partner) and ESA), cultural (UNESCO) and creative expertise (INA, CNRS, University of Leeds, IRCAM and CIANT), together with commercial partners (ACS, ASemantics, MetaWare, Engineering Ingegneria Informatica and IBM/Haifa), experts in knowledge engineering (CNR and FORTH) and academic partners working in this area (Universities of Glasgow and Urbino).

[2] http://www.lockss.org/lockss/Home

The main aims of CASPAR are:

1. to produce key infrastructure components to support digital preservation of digitally encoded information, strongly adhering to the concepts of the OAIS Reference Model (CCSDS, 2002); and
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.

Solutions or Snake-oil?

It is easy to propose some solutions – and extremely easy to make positive gestures. 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 CCLRC, 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?

The OAIS approach is essentially that there must be a way of testing that claim 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*. It 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 a Data Object which itself needs its own Representation Information. The Representation Networkwhich OAIS defines as the *set of Representation Information that fully describes the meaning of a Data Object*, is a very large collection, as the box "Representation Information for Martians" below indicates.

Representation Information for Martians

How much Representation Information would one need to provide for a Martian to understand and use Ionosonde[1] 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.

1 see http://en.wikipedia.org/wiki/Ionosonde

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 (its Knowledge Base), and

therefore does not require 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.

Validation Metrics

CASPAR proposes[3] 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 a practical demonstration by means of what may be regarded as "accelerated lifetime" tests involving:
    1. environment (including software, hardware) 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.

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 or moderating content – and which has become one of the most authoritative (or at least most Googled) sources of information on the Internet. Similar efforts 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 relies not on its own judgement of the value of a page but rather on the value others 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 be more than just the equivalent of a Wiki; there is a need to be proactive otherwise the information will be lost through neglect, and CASPAR's preservation infrastructure components are intended for just this purpose.

[3] CASPAR Description of Work http://www.casparpreserves.eu/Members/metaware/ReferenceDocuments/caspar-description-of-work/at_download/file Table 1 - Digital Preservation Metrics

[4] See for example http://www.bbc.co.uk/radio4/science/findlistenlabel/?programme=allinthemind20070410

Preservation Infrastructure

Key components in a preservation infrastructure need to facilitate the capture and use of Representation Information. Figure 1 shows a Strawman Architecture which is described in much greater detail in the CASPAR Conceptual Model[5].

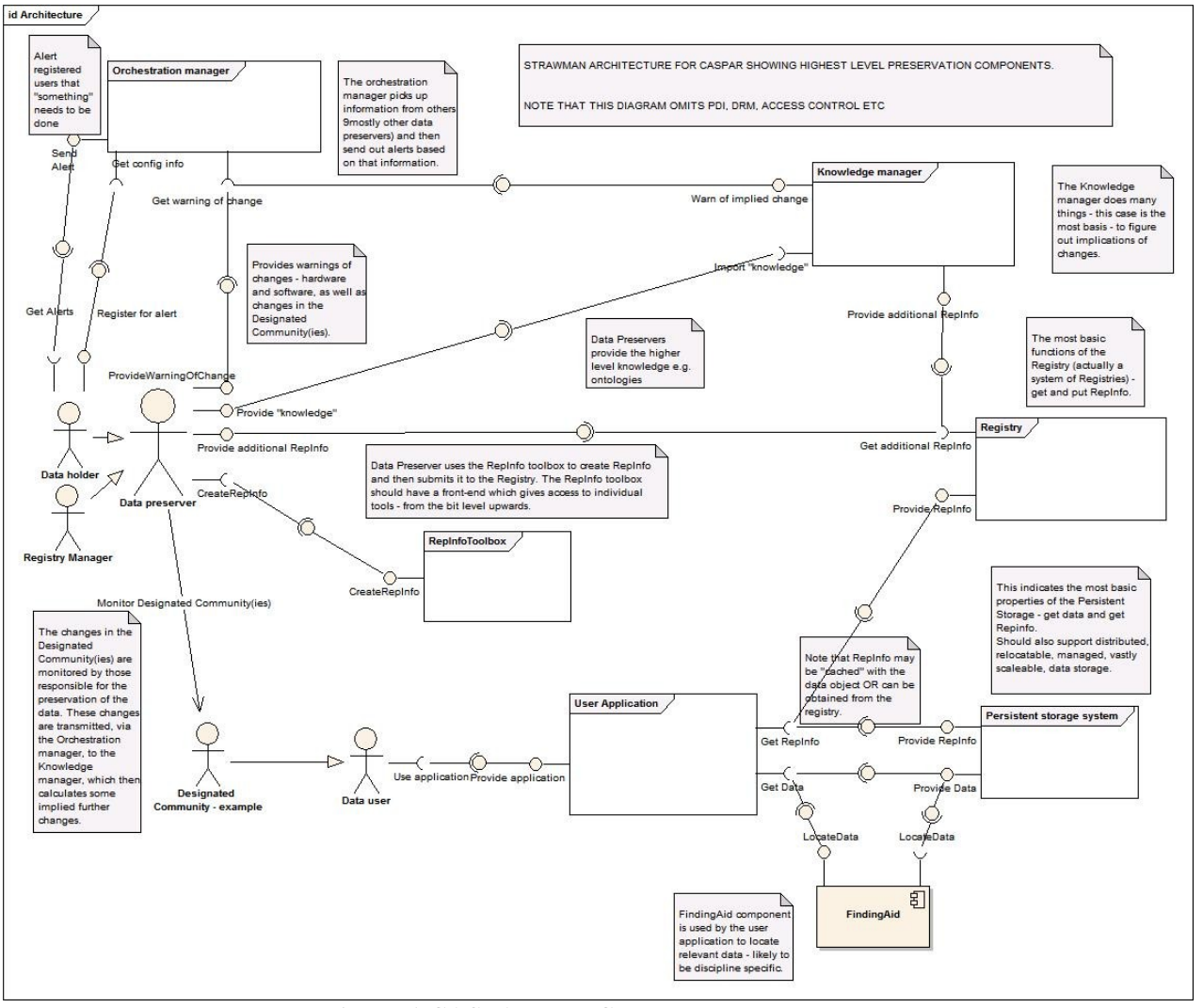

Figure 1 CASPAR Key Components

It is worth pointing out that much of the preliminary development work[6] for the Registry/Repository has been done by the Digital Curation Centre[7]. The Registry/Repository allows Representation Information to be stored, found and shared.

Identifiers (called here Curation Persistent Identifiers (CPIDs)) are associated with any data object, and point to the appropriate Representation Information in a Registry/Repository, as illustrated in Figure 2. The Representation Information returned by the Registry/Repository itself is a digital object with its own CPID.

[5] http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-conceptual-model-phase-1-1/at_download/file
[6] http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCApproachToCuration
[7] http://www.dcc.ac.uk

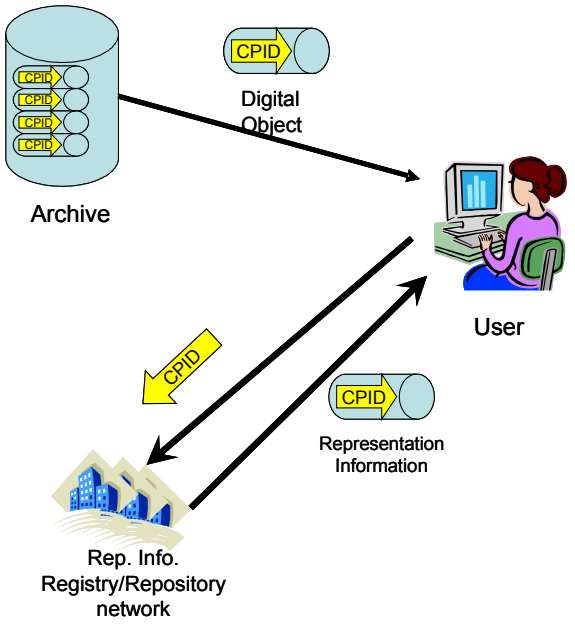

Figure 2 Use of Registry/Repository of Representation Information

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 Registry/Repository is supplemented by the Knowledge Manager – more specifically a Representation Information Gap manager which identifies gaps which need to be filled. 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 collects this information and distributes.

Support for automation in identifying such "gaps" , based on information received, is illustrated in Figure 3 (below) 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.

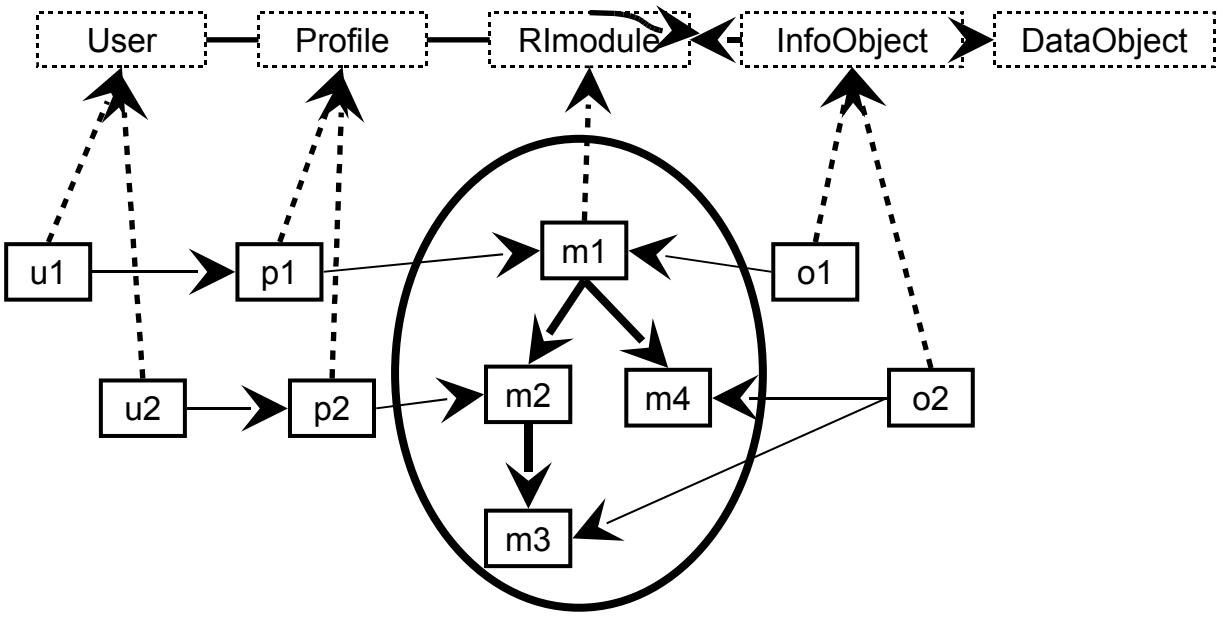

Figure 3 Modelling Users, Profiles, Modules and Dependencies

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 is available (Tzitzikas, 2007; Tzitzikas & Flouris, 2007).

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

The Warwick Workshop[8] noted that Virtualisation is an underlying theme, with a layering model illustrated in Figure 4 below.

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

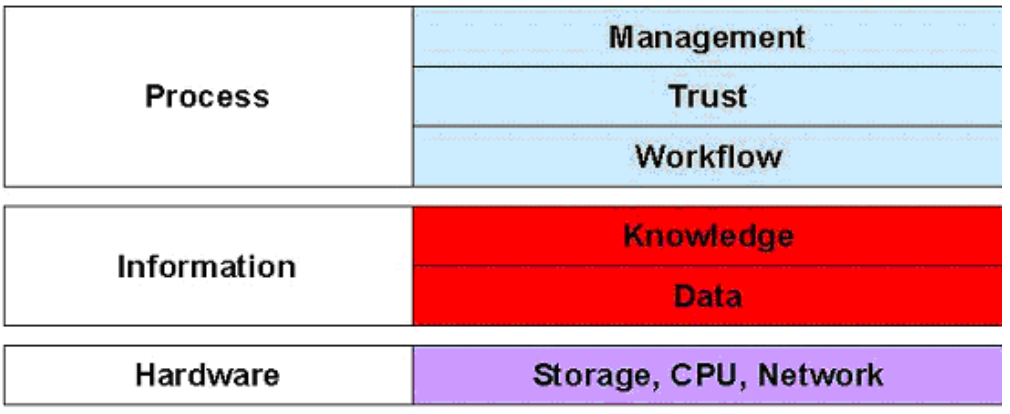

Figure 4 Layers of Virtualisation

However virtualisation is not a magic bullet. One cannot expect it to be applied everywhere, and even where it can be, the interfaces may themselves become obsolete and will eventually have to be re-engineered or 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.

The Wikipedia entry for Virtualisation[9] provides an extensive list of types of virtualisation, and distinguishes between:

- platform virtualisation, which involves the simulation of virtual machines;
- resource virtualisation, which involves the simulation of combined, fragmented, or simplified resources.

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.

[9] http://en.wikipedia.org/wiki/Virtualization

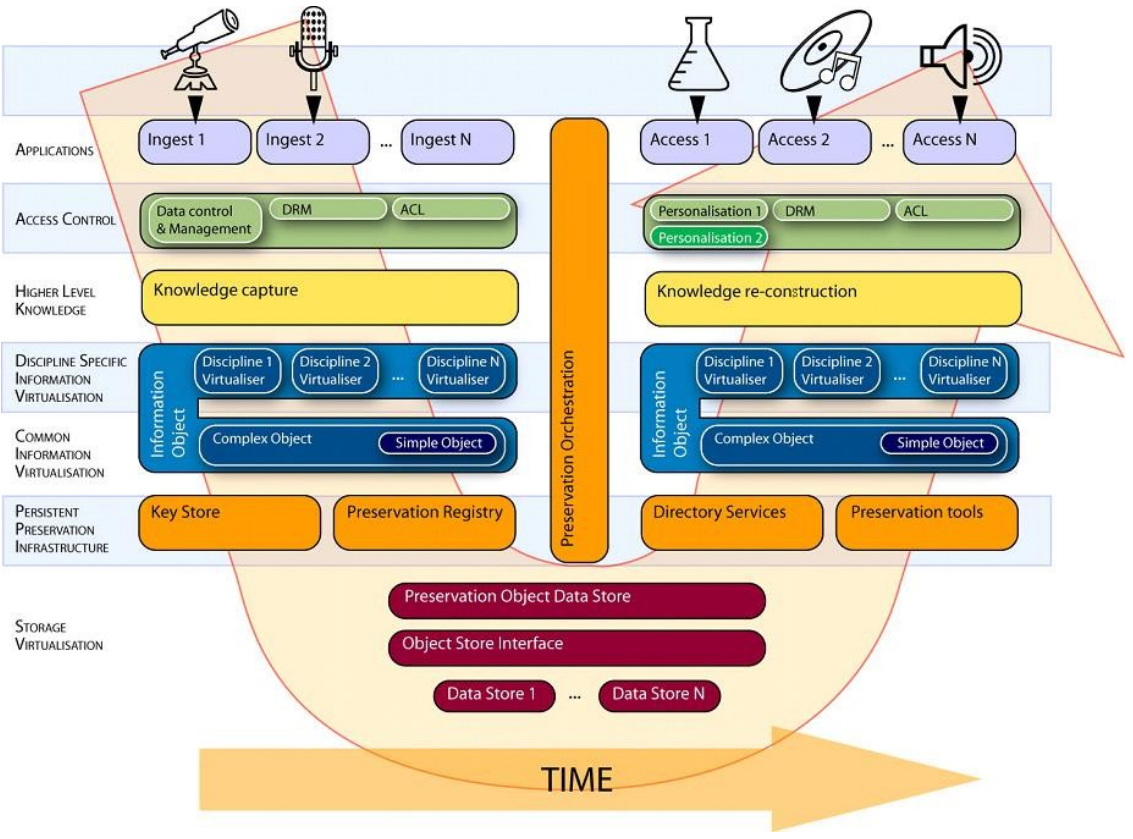

Figure 5 CASPAR Information Flow Architecture

Figure 5 indicates in somewhat more detail how CASPAR expects to use Virtualisation including:

- Digital Object Storage virtualisation
- Common information virtualisation
- Discipline-specific information virtualisation

Virtualisation also applies to the

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

Summary

CASPAR is attempting to use OAIS concepts rigorously and to the fullest extent possible, supplementing them where appropriate. 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[10].

Acknowledgements

I would like to thank the following colleagues for their contribution to this article: Luigi Fusco (ESA/ESRIN), Seamus Ross, (HATII, University of Glasgow), Mariella Guercio (University of Urbino), Mario Hernandez (UNESCO), Ugo Di Giammatteo (ACS), Zavisa Bjelogrlic (ASemantics), Dalit Naor (IBM/Haifa), Carlo Meghini

[10] http://www.casparpreserves.eu

(CNR), Silvia Boi (MetaWare), Daniel Teruggi (INA), Kia Ng (University of Leeds), Luigi Briguglio (Engineering Ingegneria Informatica S.p.A.), Vassilis Christophides (FORTH), Bruno Bachimont (CNRS/UTC), Hugues Vinet (IRCAM) and Pavel Sedlak (CIANT).

References

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

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.

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.