



Project no. 033572

CASPAR

Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval

Instrument: Information Society Technologies

Thematic Priority: 2.5.10 Access to and preservation of cultural and scientific resources

REVIEW OF STATE OF THE ART



Document identifier:	CASPAR-D1101-TN-0101-1_0
Submission Date:	15-05-2007
Due date:	30-09-2006
Work package:	1100
Partners:	All Partners
WP Lead Partner:	INA
Document status	FINAL

Abstract: This document provides a survey of the state of the art in digital preservation technologies, as a baseline and comparison for the work being done in **CASPAR**.



Delivery Type Report
 Author(s) **CASPAR Consortium**

 Approval David Giaretta
 Summary
 Keyword List
 Availability PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0_0	10 Jan 2007	Initial draft bringing together contributions	
	12 Jan 2007	Comments from Univ. Leeds	
	18 Jan 2007	Comments from CCLRC	Brian Matthews
	07 Feb 2007	Additions from CCLRC	David Giaretta
	15 Feb 2007	Interim editing and restructuring by CCLRC	Simon Lambert
	2 April 2007	Comprehensive editing and tidying by HATII and CCLRC	Jill Spellman, Simon Lambert
	16 April 2007	Additions from Metaware	Fiore Basile
	5 May 2007	Pre-final version	Jill Spellman, Simon Lambert, contributions from other partners
0_1	13 May 2007	Candidate deliverable	David Giaretta
1_0	15 May 2007	Final version, submitted deliverable	Simon Lambert





Project information

Project acronym:	CASPAR
Project full title:	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
Proposal/Contract no.:	IST-2006-033572

Project Officer: Carlos Oliveira

Address:	INFISO-E3 Information Society and Media Directorate General Content - Learning and Cultural Heritage Postal mail: Bâtiment Jean Monnet (EUFO 1167) Rue Alcide De Gasperi / L-2920 Luxembourg Office address: EUROFORUM Building - EUFO 1167 10, rue Robert Stumper / L-2557 Gasperich / Luxembourg
Phone:	+352 4301 33052
Fax:	+352 4301 33190
Mobile:	
E-mail:	Carlos.Oliveira@ec.europa.eu

Project Co-ordinator: David Giaretta

Address:	CCLRC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	d.l.giaretta@rl.ac.uk





CONTENTS

1	INTRODUCTION	7
1.1	BACKGROUND	7
1.2	PURPOSE OF THE DELIVERABLE	7
1.3	APPLICABLE DOCUMENTS AND REFERENCE DOCUMENTS.....	8
2	METHODOLOGY	9
2.1	WHAT TO CONSIDER	9
2.2	RELATION TO OAIS.....	12
2.2.1	<i>The Functional Model</i>	12
2.2.2	<i>The Information Model</i>	12
2.2.3	<i>Preservation Description Information</i>	12
2.2.4	<i>Packaging Information</i>	12
2.2.5	<i>Digital Rights Management and access control</i>	12
2.2.6	<i>Proof of preservation effectiveness</i>	12
2.3	THE QUESTIONNAIRE AND ASSESSMENT	12
2.4	STRUCTURE OF THE DELIVERABLE.....	12
3	OAIS FUNCTIONAL ENTITIES.....	12
3.1	INGEST	12
3.1.1	<i>BRICKS</i>	12
3.1.2	<i>Fedora</i>	12
3.1.3	<i>kopal</i>	12
3.1.4	<i>MUSTICA</i>	12
3.1.5	<i>PAIMAS</i>	12
3.1.6	<i>InterPARES</i>	12
3.2	ARCHIVAL STORAGE.....	12
3.2.1	<i>BRICKS</i>	12
3.2.2	<i>Chronopolis</i>	12
3.2.3	<i>DILIGENT</i>	12
3.2.4	<i>DSpace</i>	12
3.2.5	<i>Fedora</i>	12
3.2.6	<i>e-Depot and DIAS</i>	12
3.2.7	<i>The OSD standard</i>	12
3.2.8	<i>The XAM standard</i>	12
3.2.9	<i>Storage Resource Broker (SRB)</i>	12
3.2.10	<i>iRODS</i>	12
3.2.11	<i>InterPARES</i>	12
3.3	DATA MANAGEMENT	12
3.3.1	<i>BRICKS</i>	12
3.3.2	<i>DILIGENT</i>	12
3.3.3	<i>Fedora</i>	12
3.3.4	<i>iRODS</i>	12
3.4	PRESERVATION PLANNING.....	12
3.4.1	<i>BRICKS</i>	12
3.4.2	<i>DILIGENT</i>	12
3.4.3	<i>DSpace</i>	12
3.4.4	<i>Fedora</i>	12
3.4.5	<i>MUSTICA</i>	12
3.4.6	<i>Preservation planning for virtual arts</i>	12
3.4.7	<i>SAFE</i>	12
3.4.8	<i>InterPARES</i>	12
3.5	ACCESS	12
3.5.1	<i>BRICKS</i>	12
3.5.2	<i>DILIGENT</i>	12
3.5.3	<i>Fedora</i>	12





3.5.4	OpenURL.....	12
3.5.5	OAI-PMH	12
3.5.6	Semantic web standards.....	12
3.5.7	XML/XSLT driven dissemination.....	12
4	OAIS INFORMATION MODEL.....	12
4.1	INFORMATION MODELS.....	12
4.1.1	<i>Cedars</i>	12
4.1.2	<i>DILIGENT</i>	12
4.1.3	<i>SAFE</i>	12
4.1.4	<i>CIDOC Conceptual Reference Model (CRM)</i>	12
4.1.5	<i>InterPARES</i>	12
4.2	REPRESENTATION INFORMATION.....	12
4.2.1	<i>SAFE</i>	12
4.3	STRUCTURE INFORMATION	12
4.3.1	<i>BRICKS</i>	12
4.3.2	<i>DILIGENT</i>	12
4.3.3	<i>Fedora</i>	12
4.3.4	<i>SAFE</i>	12
4.3.5	<i>Data Structure Description Languages</i>	12
4.4	SEMANTIC INFORMATION	12
4.4.1	<i>BRICKS</i>	12
4.4.2	<i>Fedora</i>	12
4.4.3	<i>SAFE</i>	12
4.5	OTHER INFORMATION.....	12
4.5.1	<i>BRICKS</i>	12
4.5.2	<i>DILIGENT</i>	12
4.5.3	<i>SAFE</i>	12
4.5.4	<i>Computer Emulation and Virtualisation Technologies</i>	12
5	OAIS PRESERVATION DESCRIPTION INFORMATION (PDI).....	12
5.1	GENERAL PRESERVATION DESCRIPTION INFORMATION	12
5.1.1	<i>Cedars</i>	12
5.1.2	<i>BRICKS</i>	12
5.1.3	<i>Fedora</i>	12
5.1.4	<i>MUSTICA</i>	12
5.1.5	<i>CIDOC Conceptual Reference Model</i>	12
5.2	PDI FOR VIRTUAL ARTWORKS	12
5.2.1	<i>Archiving the Avant-Garde</i>	12
5.2.2	<i>The Database of Virtual Art</i>	12
5.3	PDI-REFERENCE.....	12
5.3.1	<i>Persistent identification with the handle system</i>	12
5.3.2	<i>Digital Object Identifiers (DOI)</i>	12
5.3.3	<i>URI etc</i>	12
5.3.4	<i>Persistent URL (PURL)</i>	12
5.3.5	<i>ARK (Archival Resource Key)</i>	12
5.3.6	<i>Life Sciences Identifiers</i>	12
5.3.7	<i>Name to Thing (N2T)</i>	12
5.4	OAIS PACKAGING	12
5.4.1	<i>The Metadata Encoding and Transmission Standard (METS)</i>	12
5.4.2	<i>XML Formatted Data Unit (XFDU)</i>	12
6	DIGITAL RIGHTS MANAGEMENT AND ACCESS CONTROLS.....	12
6.1	OVERVIEW.....	12
6.2	SECURITY IN OTHER PROJECTS.....	12
6.3	USEFUL REFERENCES.....	12
6.3.1	<i>DRM policy creation</i>	12
6.3.2	<i>DRM policy projection</i>	12





6.3.3	DRM security and cryptography	12
6.3.4	Access control (AC) technologies.....	12
7	PROOF OF PRESERVATION EFFECTIVENESS	12
7.1	BRICKS.....	12
7.2	DILIGENT.....	12
7.3	FEDORA.....	12
7.4	SAFE	12
8	SUMMARY.....	12





1 INTRODUCTION

1.1 BACKGROUND

An awareness of the need for preservation of digital assets has been growing for several years. The motivation has included fear of the decay of older physical media—such as recordings of dance performances on videotape—and awareness of the difficulties experienced in correctly interpreting scientific data after many years. In the business world, digital archive capacity is expected to significantly increase in the coming years, driven by regulatory compliance, corporate governance, desire to protect the investment put into creating the information as well as addressing risks of litigation and loss of prestige. In general, archive data covers a myriad of different record types including numerical data, digital images, office documents, scanned paper and many others, each with particular preservation needs and priorities.

A number of major initiatives have been set up in response to this awareness. Looking only at the UK, the Cedars project ran for three years from 1998 and was a coordinated effort to investigate the issues around digital preservation from a research library perspective. The CAMiLEON project is a successor to Cedars, and is a joint UK–US initiative focussing particularly on emulation as a preservation strategy.

The Digital Preservation Coalition was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure the global digital memory and knowledge base. The Digital Curation Centre was created in 2003 with the purpose of providing a national focus for research and development into curation issues and to promote expertise and good practice, both national and international, for the management of all research outputs in digital format. In this context, digital curation means maintaining and adding value to a trusted body of digital information for current and future use, which encompasses preservation.

The European RTD programmes have also recognised this trend, and already within the 5th Framework Programme there were some projects in the area, such as the ERPANET Network of Excellence. In the 6th Framework Programme came its successor DPE project, PrestoSpace on audiovisual content, and Planets concentrating on planning of preservation actions.

Currently, the benchmark standard for the construction of preservation environments is the Open Archival Information System (OAIS) Reference Model, which is an initiative of NASA's Consultative Committee for Space Data Systems. OAIS is of course the guiding principle of CASPAR.

It is against the background of this activity that the CASPAR must operate, taking account of relevant work in the area of preservation itself, as well as being aware of technologies that could be suitable as a basis for CASPAR's own developments.

1.2 PURPOSE OF THE DELIVERABLE

As in any large-scale project with far-reaching aims, the purpose of the state-of-the-art review is to inform and motivate the choices made in the project by means of a critical and reasoned review of literature and current practices. Identification of the best ideas in the various relevant areas is in particular helping to guide the development of the conceptual model, the overall architecture and component design for CASPAR.

The CASPAR Description of Work, in its description of WP1100 'Review State of the Art', notes that in the different fields covered by CASPAR 'there has been increasing activity in these domains during recent years and proposals as well as initiatives are flourishing in many countries or international bodies.' The work package aims at 'providing clear cartography of different ongoing projects and at focusing on the main issues that remain critical for the development of long-term preservation





solutions. Another crucial goal for this WP is to provide a common understanding of the state of the art within the **CASPAR** consortium in order to have a shared intellectual basis.’

The task breakdown of the work package distinguishes three relevant areas that need to be covered:

- **Methodology:** preservation essentially consists in managing a process and organising data and knowledge.
- **Research and technological solutions:** as digital preservation becomes recognised as a pervasive problem, industrial solutions for parts of the problem are becoming available.
- **Existing standards:** standards are obviously an important way to obtain interoperability and exchangeability of data, one of the issues of preservation.

When considering the state of the art, an important distinction is between *positioning* and *selection*. Positioning consists in understanding the relationship of **CASPAR** with respect to other initiatives; it might or might not entail building on their concrete results, but it means being aware of what the differences are, or the advances that **CASPAR** will bring, or accepting not to work in a certain area if that has already been adequately covered. Selection refers to the choice of methods and tools that are required for **CASPAR**’s own developments, whether they originate with preservation-related initiatives or simply because they are valuable and contribute to the achievement of the project’s goals. An example of the latter is Semantic Web technologies and certain storage technologies: these are not specifically preservation-oriented, but have something useful to offer the project.

The original expectation was that this deliverable would be produced quite early in the project so as to serve as a secure foundation for the work that follows. However it has been found to be more realistic for the state of the art review to proceed partly in parallel with the initial stages of the architecture and modelling work. Technological changes and research carry on apace. This document is a snapshot of current ideas. It is planned to continue to keep informed about development in the state of the art throughout the life of the project. For example, preservation-related European projects running in parallel with **CASPAR**, such as Planets, plan to produce outputs that will probably be of relevance and should be considered in due course.

It should be noted that the work of the Digital Curation Centre underlies the **CASPAR** approach and so has not been explicitly considered in this review of the state of the art—it is more like a baseline for the project. For example, the Representation Information Repository forms an important part of the **CASPAR** framework. This was always envisaged: the original **CASPAR** proposal noted that ‘Initial studies of such a Registry have been undertaken by the UK Digital Curation Centre (DCC). ... This work will be pursued and enlarged by **CASPAR**.’

1.3 APPLICABLE DOCUMENTS AND REFERENCE DOCUMENTS

Applicable documents

[A1] Description of Work, April 2006

[A2] Risk Form

Reference documents

[R1] **CASPAR** project proposal, Sept 2005

Glossary

Please note that a glossary is available in Annex I – Glossary.

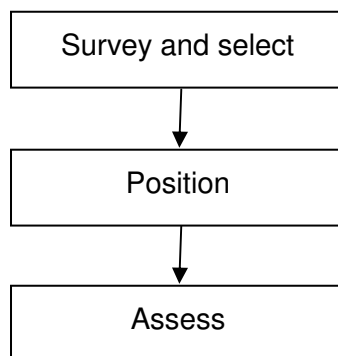




2 METHODOLOGY

2.1 WHAT TO CONSIDER

The methodology that has been adopted is in three stages.



The first stage, ‘Survey and select’, consists in taking a wide-ranging view of the landscape and picking out the most important developments that together cover the areas of concern of CASPAR. It is not necessary to identify every single project, initiative or standard that impinges in any way on digital preservation; it is sufficient to choose those that are considered by the community (or communities in particular domains such as performing arts) to be of especial importance and currency. These will include major national and international projects, influential technologies and *de facto* standards for particular domains.

The CASPAR partners have worked together to produce such a list, and it forms the basis of the whole deliverable. In summary form, the selected items are shown in the following table. The table is organised into blocks, corresponding to the nature of the connection with digital preservation:

- a specific focus on preservation;
- preservation by implication (i.e. preservation is not the main focus, but arises incidentally);
- a potential tool or element of a preservation environment;
- other relevant bases for CASPAR’s work.

In framing the review of the state of the art, there is an expectation that the reader will be familiar with current Information Technology concepts and methods. Those concepts or methods that the reader could expect to find in a standard text book or even wikipedia have been assumed to be knowledge dependencies that readers could be expected to meet.

Name	Short summary	Type	Domain
Specific focus on preservation			
Archiving the Avant-Garde	Documenting and preserving digital / variable media art. It has two case studies/related project ‘Renewing the Erl King’ and ‘Preserving the Rhizome ArtBase’ http://www.bampfa.berkeley.edu/about/avantgarde	Project	Digital/ variable media art
Cedars	UK project, now ended, exploring digital preservation issues including acquiring digital objects, their long-term retention, sufficient description, and eventual access. http://www.leeds.ac.uk/cedars/	Project	—
Chronopolis	NSF-funded project, aiming to be national centre for the	Project	Science





	management, long-term preservation, and promulgation of digital assets. http://globalstor.org/pdf/presentations/Moore-chronopolis.pdf		
Digital Video Preservation Reformatting Project	Project focusing on the critical need for preservation initiatives for dance recorded on videotape. http://www.danceheritage.org/preservation/	Project	Performing arts (dance)
InterPARES	A US–Canadian project, now ended, that worked on the long-term preservation of records created and/or maintained in digital form, with emphasis on authenticity. http://www.interpares.org/	Project	—
kopal	German project building a cooperatively developed and operated archival system to ensure the long-term access to digital documents. http://kopal.langzeitarchivierung.de/downloads/kopal_Broschur_e_2006_en.pdf	Project	—
MUSTICA	A preservation tool for electro-acoustic music. http://polaris.gseis.ucla.edu/blanchette/MUSTICA.html	Project, preservation tool	Electro-acoustic music
Planets	A European project bringing together National Libraries and Archives, leading research institutions, and technology companies to address the challenge of preserving access to digital cultural and scientific knowledge. http://www.planets-project.eu/	Project	—
Preserv	UK project investigating and developing infrastructural digital preservation services for institutional repositories. http://preserv.eprints.org/	Project	—
PrestoSpace	European project on preservation towards storage and access: standardised practices for audio-visual content. http://cordis.europa.eu/ist/digicult/presto.htm	Project	Audio-visual content
Variable Media Network	Proposes an unconventional new preservation strategy that has emerged from the Guggenheim's efforts to preserve its world-renowned collection of conceptual, minimalist and video art. http://www.variablemedia.net/e/welcome.html	Project and network of organisations	Art in new and ephemeral media
Preservation by implication			
BRICKS	European project with worldwide community, researching and implementing advanced open source software solutions for the sharing and exploitation of digital cultural resources. http://www.brickcommunity.org/	Project	Cultural heritage
DART	Proof-of-concept project to develop tools to support the new collaborative research infrastructure of the future. http://www.dart.edu.au/	Project	—
Database of Virtual Art	Documents the rapidly evolving field of digital installation art, with a view to systematic preservation of this art in future. http://www.virtualart.at/	Project, database	Digital installation art
DELOS	European Network of Excellence on Digital Libraries. http://www.delos.info/	Network of organisations	—
DILIGENT	Project integrating Grid and digital library technologies, creating an advanced test-bed that will allow virtual e-Science	Project	Science





	communities to share knowledge and collaborate in a secure, coordinated, dynamic and cost-effective way. http://www.diligentproject.org/		
DSpace	Digital repository system that captures, stores, indexes, preserves, and distributes digital research material. http://www.dspace.org/	Software system	—
e-Depot, DIAS	Digital archiving system of the National Library of the Netherlands. http://www.kb.nl/e-depot	System	—
Fedora	Open source software for creating flexible service-oriented architecture for managing and delivering digital content. http://www.fedora.info/	Software system	—
Tool or element of preservation environment			
OSD, XAM	Standards enabling the encapsulation of large metadata with raw data for long-lived information. http://www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf	Standard	—
PAIMAS	Producer-Archive Interface Methodology Abstract Standard http://public.ccsds.org/publications/archive/651x0b1.pdf	Standard	—
SAFE	The SAFE (Standard Archive Format for Europe) has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. http://earth.esa.int/SAFE/	Standard	Earth observation data
SRB, iRODS	Systems for managing distributed and persistent data. http://www.sdsc.edu/srb/index.php/Main_Page , http://irods.sdsc.edu/index.php/Main_Page	Software system	—
Universal Virtual Computer	A simple virtual computer architecture that will run on any existing hardware platform. http://www.rlg.org/legacy/preserv/diginews/diginews5-3.html	Standard	—
XFDU	XML Formatted Data Unit is a standard for facilitating information transfer and archiving and is a technique for packaging data objects and representation information into a single packaged unit. http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206610R1/Attachments/661x0r1.pdf	Standard	—
Other relevant bases			
CIDOC CRM	Formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. http://cidoc.ics.forth.gr/	Standard	Cultural heritage
Data Structure Description Languages	Languages expressing mapping data bits within a data stream/file to data values such as numbers and strings and then showing how these values are ordered with respect to one another within the data hierarchy.	Standard	—
DOI, handle	Approaches to persistent identification of resources. http://www.doi.org/	Standard, software system	—
OAI-PMH	Protocol for Metadata Harvesting (OAI-PMH), providing a standard way for a data repository to expose metadata to third	Standard	—





	party indexing engines. www.openarchives.org/OAI/openarchivesprotocol.html		
METS	Metadata Encoding and Transmission Standard is an open extensible XML schema standard developed by the library community providing the means to convey Representation Information necessary for the management of digital objects within an archive or repository and the exchange of objects between repositories or repositories and consumers. http://www.loc.gov/standards/mets/	Standard	—
OpenURL	A type of URL that contains resource metadata for use primarily in libraries. http://www.oclc.org/research/projects/openurl/default.htm	Standard	—
Provenance	Project, now ended, to conceive, design and implement an industrial strength open provenance architecture for grid systems. http://twiki.gridprovenance.org/bin/view/Provenance/	Project	—
Semantic web standards	Set of standards such as RDF and OWL arising from the semantic web community. http://www.w3.org/2000/01/sw/	Standard	—
XML, XSLT	Web standards http://www.w3.org/XML/ , http://www.w3.org/TR/1999/REC-xslt-19991116	Standard	—

Table 1 Summary of projects and other initiatives considered

2.2 RELATION TO OAIS

The second stage of the methodology, ‘Position’, involves relating the selected items to the **OAIS** Reference Model. As this model is fundamental to **CASPAR**, it is the ideal framework on which to hang the assessment of the state of the art. A number of elements of the model are employed.

2.2.1 The Functional Model

The **OAIS** Functional Model is illustrated in Figure 1.

A brief description of each of the entities is as follows.

Ingest. This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive’s data formatting and documentation standards, extracting Descriptive Information from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.

Archival Storage. This entity provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders.



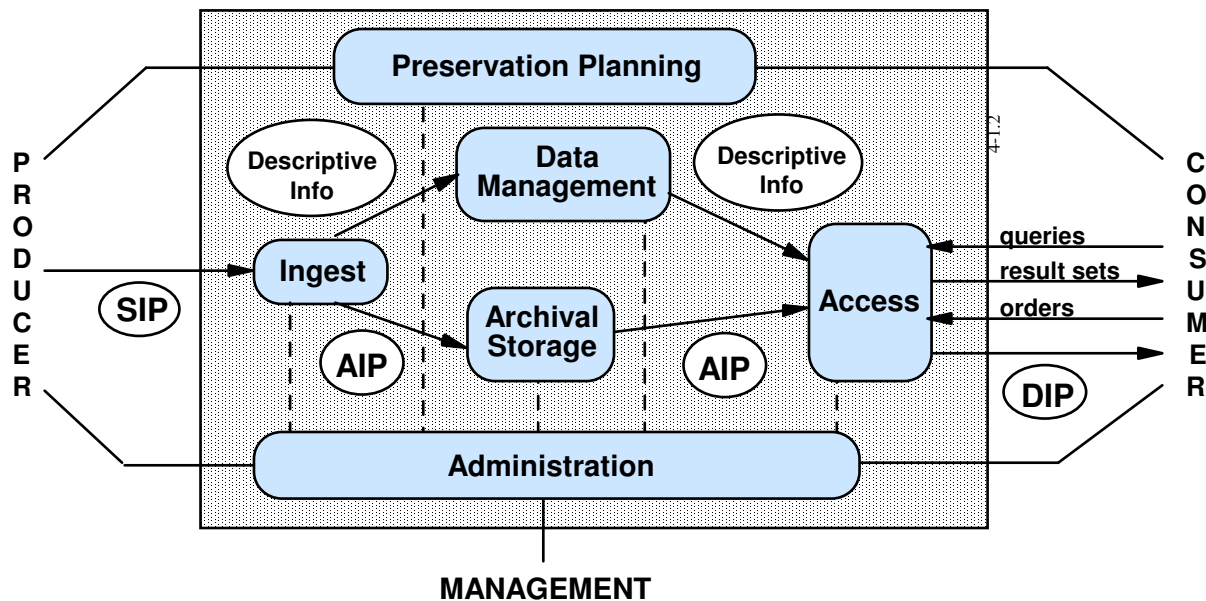


Figure 1 The OAIS Functional Model

Data Management. This entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents archive holdings and administrative data used to manage the archive. Data Management functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data), performing queries on the data management data to generate result sets, and producing reports from these result sets.

Administration. This entity provides the services and functions for the overall operation of the archive system. Administration functions include soliciting and negotiating submission agreements with Producers, auditing submissions to ensure that they meet archive standards, and maintaining configuration management of system hardware and software. It also provides system engineering functions to monitor and improve archive operations, and to inventory, report on, and migrate/update the contents of the archive. It is also responsible for establishing and maintaining archive standards and policies, providing customer support, and activating stored requests.

Preservation Planning. This entity provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base. Preservation Planning also designs IP templates and provides design assistance and review to specialize these templates into SIPs and AIPs for specific submissions. Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

Access. This entity provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products. Access functions include communicating with Consumers to receive requests, applying controls to limit access to specially protected information, coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to Consumers.





In addition to the entities described above, there are various **Common Services** assumed to be available. These services are considered to constitute another functional entity in this model. This entity is so pervasive that, for clarity, it is not shown in the figure.

It should be noted that Administration is not included in the framework for positioning of the state of the art. This is because its implementation tends to be a human-level process and highly specific to particular archives, and so less susceptible to general analysis. Neither are Common Services covered: these are at the opposite extreme, low-level technical services such as inter-process intercommunication and exception handling.

2.2.2 The Information Model

As well as these functional entities, the **OAIS** Information Model is also employed as reference. A basic concept of the **OAIS** Reference Model (ISO 14721) is the concept of information being a combination of data and Representation Information. The Unified Modeling Language (UML) diagram in Figure 2 illustrates this concept. The Information Object is composed of a Data Object that is either physical or digital, and the Representation Information that allows for the full interpretation of the data into meaningful information. This model is valid for all the types of information in an **OAIS**.

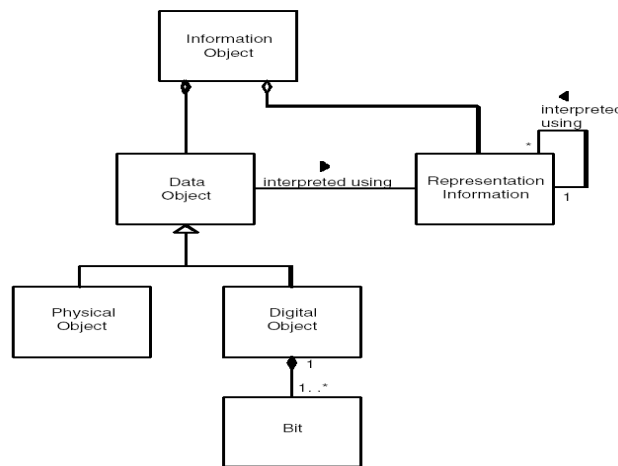


Figure 2 OAIS Information Model

This UML diagram shows 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;¹
- 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.

This figure shows that Representation Information may contain references to other Representation Information. Representation Information is an Information Object that may have its own Digital Object and other Representation Information associated with understanding each Digital Object, as shown in a compact form by the interpreted using association, the resulting set of objects can be referred to as a Representation Network.

¹ An example of the former is a piece of paper or a rock sample.





The Representation component Figure 3 shows more details and in particular breaks out the semantic and structural information as well as recognizing that there may be “Other” representation information such as software.

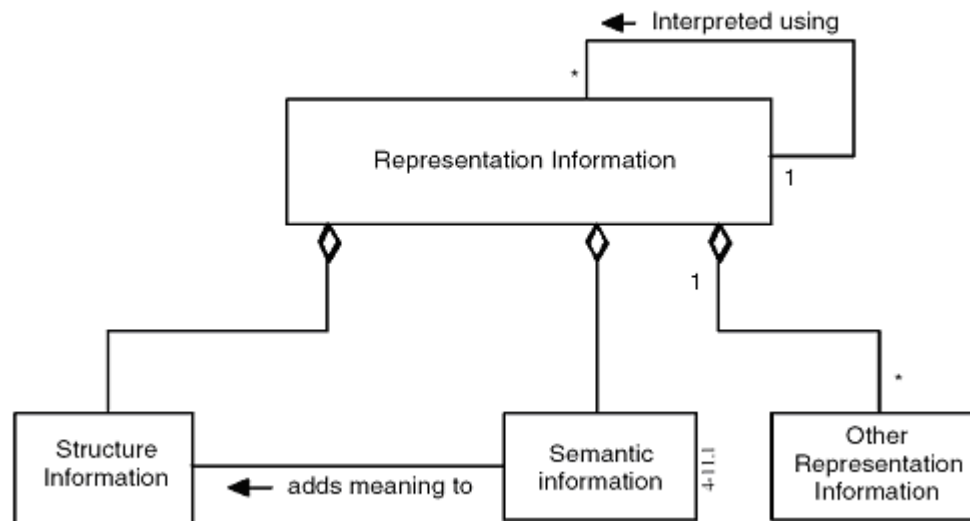


Figure 3 Representation Information Object

The recursion of the Representation Information will ultimately stop at a physical object such as a printed document (ISO standard, informal standard, notes, publications, etc) and the use of paper documentation tends to prevent ‘automated use’ and ‘interoperability’. Complete resolution of the complete Representation Net to this level would be an almost impossible task. Therefore we would prefer to stop earlier. In particular we can stop for a particular designated community when the Representation Information can be understood with that designated community’s Knowledge Base.

A science file in Flexible Image Transport System (FITS) format will be understood by someone who knows how to handle this format (whose Knowledge Base includes FITS) and has appropriate software. Someone whose Knowledge Base does not include FITS will need additional Representation Information and software or the written FITS standard.

A problem with Representation Information is that the amount needed for a particular object could be vast and impractical to do anything with in reality. It is for that reason that the concept of the designated community is so important. It allows us to limit the Representation Information required to be captured at any one time, and track what needs to be changed in the Representation Information as the designated community changes.

2.2.3 Preservation Description Information

In the OAIS model, an Information Package is a conceptual container of two types of information called Content Information and Preservation Description Information (PDI). The PDI is the information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context Information.

2.2.4 Packaging Information

The Packaging Information is that information which, either actually or logically, binds, identifies or relates the Content Information and PDI.



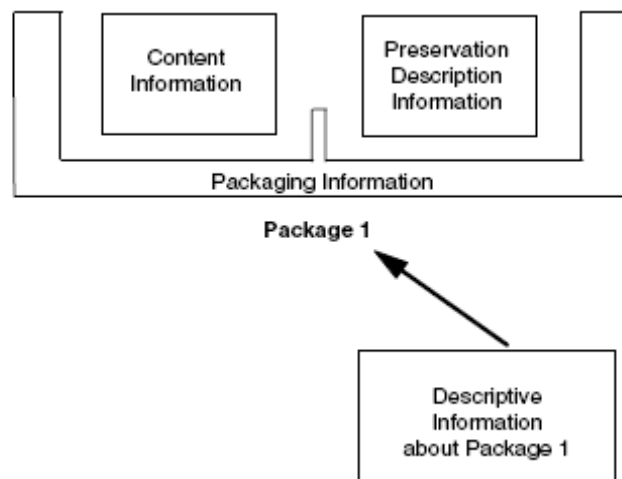


Figure 4 OAIS Information Package concepts and relationships

- **Archival Information Collection (AIC):** Archival Information Package whose Content Information is an aggregation of other Archival Information Packages.
- **Archival Information Package (AIP):** An Information Package, consisting of the Content Information and the associated PDI, which is preserved within an **OAIS**.
- **Archival Information Unit (AIU):** An Archival Information Package whose Content Information is not further broken down into other Content Information components, each of which has its own complete PDI. It can be viewed as an atomic AIP. An example of an AIU would be a table of numbers representing temperatures in a certain region with all the associated documentation describing how and where the temperatures were measured, what instruments were used to make the measurements, who made the measurements, why they were made, what processing has been performed on the measurements and who has had custody of these measurements since they were first created, how the measurements relate to other information, how the measurements can be uniquely referenced by others, etc.
- **Descriptive Information:** The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of **OAIS** information holdings by Consumers.
- **Dissemination Information Package (DIP):** The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the **OAIS**.
- **Information Package:** The Content Information and associated PDI, which is needed to aid in the preservation of the Content Information. The Information Package has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information.
- **Package Description:** The information intended for use by Access Aids.
- **Packaging Information:** The information that is used to bind and identify the components of an Information Package. For example, it may be the ISO 9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information.
- **Submission Information Package (SIP):** An Information Package that is delivered by the Producer to the **OAIS** for use in the construction of one or more AIPs.





2.2.5 Digital Rights Management and access control

OAIS has little to say about DRM. Most current DRM systems are mainly focused on proprietary mechanisms for copy protection of digital artefacts. Nonetheless it is clear that DRM and access control must be taken into account in a full preservation environment. Thus it is necessary to review the DRM approaches of other projects and initiatives and to identify their strengths and weaknesses.

2.2.6 Proof of preservation effectiveness

No matter how impressive the array of methods and tools put in place, the only touchstone of success is the effectiveness in preserving real digital assets. There must be a clear case why a particular preservation environment is likely to achieve this over a possibly very long timescale. Sometimes semi-empirical tests such as accelerated lifetime tests can be performed; otherwise arguments based on openness and extensibility may be employed.





The items selected in the first stage can now be positioned according to their relation to the framework elements just outlined.

	Ingest	Archival Storage	Data Management	Preservation Planning	Access	Information Models	Preservation Description Information	Packaging	DRM / Access Control	Proof of Pres. Effectiveness
Specific focus on preservation										
Archiving the Avant-Garde										
Cedars										
Chronopolis										
Digital Video Preservation Reformatting Project										
InterPARES										
kopal										
MUSTICA										
Planets										
Preserv										
PrestoSpace	Covered separately in annex									
Variable Media Network										
Preservation by implication										
BRICKS										
DART										
Database of Virtual Art										
DELOS										
DILIGENT										
DSpace										
e-Depot, DIAS										
Fedora										





	Ingest	Archival Storage	Data Management	Preservation Planning	Access	Information Models	Preservation Description Information	Packaging	DRM / Access Control	Proof of Pres. Effectiveness
Tool or element of preservation environment										
OSD, XAM										
PAIMAS										
SAFE										
SRB, iRODS										
Universal Virtual Computer										
XFDU										
Other relevant bases										
CIDOC CRM										
Data Structure Description Languages										
DOI, handle										
OAI-PMH										
METS										
OpenURL										
Provenance										
Semantic web standards										
XML, XSLT										

Table 2 Relevance of state of the art items





2.3 THE QUESTIONNAIRE AND ASSESSMENT

The final stage of the methodology, 'Assess', has been based on the use of a structured questionnaire applied to each of the items identified in the first stage, with a view to teasing out their relevance to **CASPAR** within the framework of **OAIS**. The questionnaire (see Annex II – The **CASPAR** State-of-the-Art Questionnaire) is in fact based on the same elements of the **OAIS** Reference Model as were used in the previous stage for positioning, though it has also been extended both in depth and in coverage (for example, it has a section on storage standards). The questionnaire has the following structure (top level only).

1. Project details
2. **OAIS** coverage
3. **OAIS** Information Model: Representation Information
4. **OAIS** Preservation Description Information
5. **OAIS** packaging
6. Preservation effectiveness
7. Trustworthiness
8. Evaluation
9. Virtualisation
10. Impact
11. DRM policy creation
12. DRM policy protection
13. DRM security and cryptography
14. Access control technologies
15. Storage standards
16. Other comments

The aim of the questionnaire is to draw out the approach and relationship of the project/initiative in question to **OAIS**, thereby allowing for useful comparisons and judgements.

Because of the diversity in scope and scale of the projects and other activities considered in the survey, the questionnaire was not applied systematically in all cases. In some cases, it was used more as a guideline for identifying what was of relevance to **CASPAR**, without necessarily working through in its entirety. In a few cases, where it was felt that the initiative in question was not sufficiently aligned with the **OAIS** approach, the questionnaire was not used, but instead the material is provided as an annex to this deliverable.





2.4 STRUCTURE OF THE DELIVERABLE

The remainder of the deliverable follows the framework for positioning the items in the state of the art. The following sections are organised on the basis described in section 2.2:

- **OAIS** functional entities
- Elements of the **OAIS** information model
- Preservation Description Information
- Packaging
- DRM and access controls
- Proof of preservation effectiveness

Each of these top-level headings is broken down to the appropriate level of detail, and under these headings appears the list of items relating to it, as illustrated in Figure 5.

Particular OAIS entity	
3	OAIS FUNCTIONAL ENTITIES21
3.1	INGEST 21
3.1.1	BRICKS 21
3.1.2	Fedora 22
3.1.3	kopal 22
3.1.4	MUSTICA 24
3.1.5	PAIMAS 25
3.2	ARCHIVAL STORAGE 27
3.2.1	BRICKS 27
3.2.2	Chronopolis 27
3.2.3	DILIGENT 28
3.2.4	DSpace 28
3.2.5	Fedora 29
3.2.6	e-Depot and DIAS 29
3.2.7	The OSD standard 30
3.2.8	The XAM standard 32
3.2.9	Storage Resource Broker (SRB) 33
3.2.10	iRODS 34
3.3	DATA MANAGEMENT 35

Projects etc. discussed with reference to entity, corresponding to entries in Table 2.

Figure 5 Illustration of the structure of the deliverable

Every project or activity is covered under those headings about which it has something to say, that is, those identified in Table 2. Thus a wide-ranging project such as DILIGENT will appear in many sections, whereas others will only appear in one.

References supporting particular points are given as footnotes. The basic reference for each item considered is the one given above in Table 1.

A number of annexes are provided for supplementary material which it was not appropriate to include in the main body of the deliverable. The reasons for this are:

- The material is a survey of a particular specialised area of preservation (for example: interactive multimedia performances).





- The material is a general overview of preservation in a particular domain (for example: the cultural heritage domain).
- The material relates to a particular project, which is either not suitable for incorporating into the structure of the main body, or is expanded in the annex (for example: PrestoSpace, where it was considered that the mapping to OAIS was not strong enough to be usable in the main body of the deliverable).
- The material deals with a particular technological approach rather than specific methods, tools or projects (for example: semantic web technologies).

The annexes are listed below.

Annex I • Glossary

Annex II • The CASPAR State of the Art Questionnaire

Annex III • Preservation of Interactive Multimedia Performances and 3D Motion Data Representation

Annex IV • Documenting cultural heritage

Annex V • DSpace and CASPAR

Annex VI • The MUSTICA Project

Annex VII • The PrestoSpace Project

Annex VIII • The European Space Agency's Multi-Mission Facility Infrastructure

Annex IX • Computer Emulation and Virtualisation Technologies

Annex X • Semantic Web Knowledge Management Components

Annex XI • RDF and OWL

Annex XII • In-Depth Analysis of Digital Rights Management

Annex XIII • Assessing Authenticity in Preserving Digital Resources





3 OAIS FUNCTIONAL ENTITIES

3.1 INGEST

(A) PROJECTS AND OTHER MAJOR INITIATIVES

3.1.1 BRICKS

Although Building Resources for Integrated Cultural Knowledge Services (BRICKS)² does not support the full Ingest function there is an established informal process for receiving SIP into BRICKS Framework nodes. SIPs can be harvested from existing archives once an Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) schedule is set up. This schedule defines targets, targets metadata schema mappings and provides options for referencing or copying content. Once this procedure is setup, no human intervention is required and it can be scheduled at varying intervals.

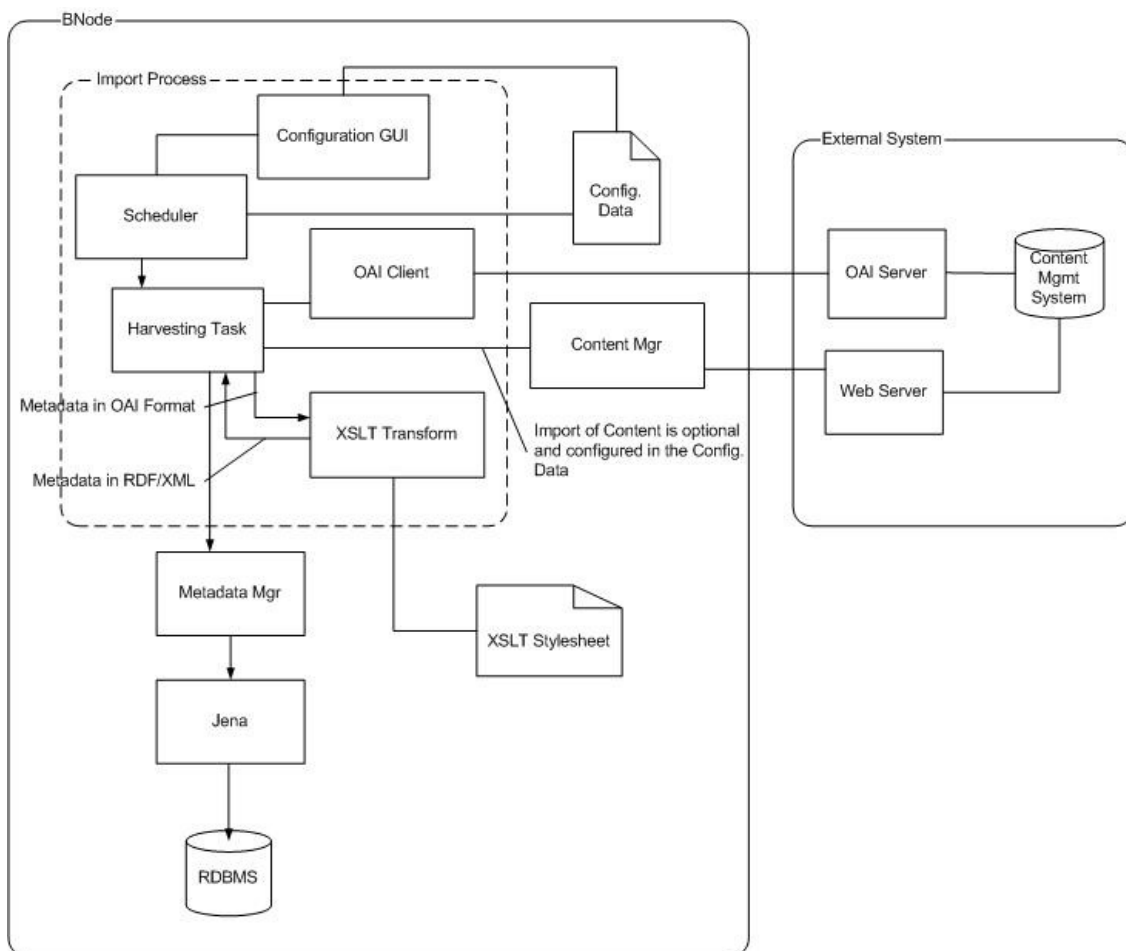


Figure 6 BRICKS: Importing Metadata (and Content) - outline of components

² <http://www.brickscommunity.org/>





3.1.2 Fedora

The Fedora Project's³ claim to **OAIS** compliance centres on its Ingest and Access functions. The following Fedora document outlines its Ingest process and compliance with **OAIS**:

<http://dca.tufts.edu/features/nhprc/reports/ingest/index.html>

Additional information about Fedora's ingest process can be found in:

<http://www.fedora.info/download/2.1.1/userdocs/digitalobjects/ingestExport.html>

3.1.3 kopal

The kopal project began in July 2004 with the aim to develop a technological and organizational solution to ensure the long-term availability of electronic publications for future generations. It pursues the building of a cooperatively developed and operated archival system to ensure long-term access to digital documents. Data must not only be physically preserved, but must be interpretable without error in the future.⁴

kopal's archival system offers a technical and organizational infrastructure with which memory organizations such as archives, libraries, and museums can make their digital collections available over the long-term. The German National Library and the Goettingen State and University Library are developing a software package for the use of the kopal solution: the "kopal Library for Retrieval and Ingest" (koLibRI). The kopal tools support the import of objects into Digital Information Archiving System (DIAS) as well as access to the archived objects.

The system is implemented according to international standards for long-term archiving and metadata within the **OAIS** framework. As such, the interface for data import (ingest) fulfills the following requirements:

- Ingest can be automated.
- The Universal Object Format (kopal-UOF) is supported.
- The flexible interface enables integration into various environments and information systems.
- International standards are utilized.
- Reuse by third parties is ensured.
- A graphical interface is in development.

Since the requirements for data export (access) are different for each institution, generic modules are used which can extend as needed, thus making possible the reuse for others. In the future, this software will also support the administration of the kopal system. Because of the quite different and in part heterogeneous system structures of the two institutions, flexible software is required.

A Workflow Tool offers a jointly usable infrastructure for modules. This Workflow Tool can serve as an asset builder for the production of archive packages. Furthermore, it can be used as a central relay to the DIAS system, as a client loader, in which it collects archive packages from many asset builders and delivers them to DIAS. Further possible uses can be integrated without problem.

The Workflow Tool follows these steps:

- **Selection:** The institution selects digital objects to be long-term archived.
- **Collection and production of metadata:** To enable the systematic storage and retrieval of objects, supplementary information such as bibliographic data is added. Technical metadata are required to be able to regularly refresh and migrate objects. The metadata are in part retrieved from information systems that may index the object in advance, and are in part generated from the object itself using special software. . Flexible data import and export

³ <http://www.fedora.info/>

⁴ . kopal Brochure, http://kopal.langzeitarchivierung.de/downloads/kopal_Broschure_2006_en.pdf





functions based on the object description scheme METS (Metadata Encoding and Transmission Standard).

- **Production of a Submission Package:** The digital objects are bundled along with their metadata as a package in a special format, the Universal Object Format (UOF).
- **Import into the long-term archive (ingest):** The software verifies the data for completeness and formal correctness before ingesting it into the archive.
- **Transformation into an archiving package:** The metadata are transferred to the data management system. The digital document and the corresponding metadata file are moved into mass storage administered by DIAS.
- **Request and retrieval of information (access):** The metadata and thus the archive packages can be accessed via the data management system.
- **Delivery:** As needed, the metadata and/or the archive material itself are delivered in a corresponding package format.
- **Utilization of the data:** Users of digital objects usually access the data via an information system. Through such a system, the user can be notified that s/he has just accessed long-term archived data. At the same time, s/he may be given the option to choose data in a particular, not necessarily current, format.

The German National Library and the Goettingen State and University Library are developing powerful kopal tools, which are precisely matched with DIAS-Core. With the completion of the project estimated for June 2007, a reusable archival system and a first release of the fully developed koLibRI-Software will be available under an Open Source license.

3.1.4 MUSTICA

The MUSTICA project⁵ (MUSTICA) was part of International Research on Permanent Authentic Records in Electronic Systems II (InterPARES), as a test case for contemporary music work archival systems. Involved partners were Groupe de Recherches Musicales of the Institut National de l'Audiovisuel (INA-GRM), Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Technologie de Compiègne (UTC) and University of California Los Angeles (UCLA). For more details see Annex VI – The MUSTICA Project.

The project provides a data model and a musical work archival system implementing that model. The project ended in 2003 and no maintenance was foreseen. As a result it should be considered still in its research stage and far from being an OAIS compliant system. However it is progress compared to current musical work preservation practices, and provides some interesting approaches to resolve the complex and heterogeneous nature of musical works.

The MUSTICA data model is formalized as a set of eXtensible Markup Language (XML) Schemas, which means that any MUSTICA description can be encoded in XML and can be validated against the MUSTICA XML Schema.

From an OAIS point of view, the Ingest process of MUSTICA is unsatisfying. Data accuracy and completeness validation beyond what the XML Schemas are able to model is restricted to the Producer. This is difficult to resolve. Musical works are extremely heterogeneous objects. The purpose of a contemporary musical work is seldom to transmit a given audio content to the listener. It is more likely something that will create an event or situation that plays with interactions between executors, electronics, public, etc. There is the potential that only the final output would generate audio content, audio content that may differ performance to performance. This is why a recording of

⁵ <http://polaris.gseis.ucla.edu/blanchette/MUSTICA.html>





an execution of the work is usually not sufficient and why the composer appears to be the only one who can make correct judgements about what is and is not important.

A more effective validation process would be (1) to attempt to re-perform the work without any assistance from the composer to verify that this is possible and (2) to ensure that the end result does not violate the authenticity of the original work. This requires an enormous effort and may be unrealistic in most contexts.

The main work flow starts with a (registered) composer who uploads onto the MUSTICA File Transfer Protocol (FTP) server a directory containing all the audio files, application data, score files, technical schemes, etc., needed to perform a musical work. After this step, the composer will:

- create a new work through the administration user interface via the web browser;
- create the first version; attach the uploaded archive(s) to it;
- fully describe everything, down to the single files stored in the archive. Once this is done, the composer will publish the work, which will become accessible (with limitations set by the composer) for the consumers on the MUSTICA access user interface.

Another scenario is if the composer (or somebody else with the permission of the composer) has ported the work to other hardware, software, or for a different performance environment. In this case a new version of the work exists and can be registered in MUSTICA using the administration interface.

3.1.5 PAIMAS

The Producer Archive Ingest Methodology Abstract Standard (PAIMAS) (ISO 20652) seeks to identify, define and provide structure to the relationships and interactions between an information Producer and an Archive. It defines the methodology for the structure of actions that are required from the initial time of contact between the Producer and the Archive until the objects of information are received and validated by the Archive.

These actions cover the first stage of the Ingest process as defined in the **OAIS** Reference Model. This recommendation describes parts of the functional entities: Administration ('Negotiate Submission Agreement') and Ingest ('Receive Submission' and 'Quality Assurance').

The standard:

- identifies the different phases in the process of transferring information between a Producer and an Archive;
- defines the objective of each phase, the actions that must be carried out during the phase, and the expected results (i.e. administrative, technical, contractual) at the end of a phase;
- forms a general methodological framework, which should be able to be applied and reused in those processes that relate to the Producer-**OAIS** Archive interface (this general framework should also provide sufficient flexibility for each particular case);
- forms a basis for the identification and/or development of standards and implementation guides in the designated community in question;
- forms a basis for identification and/or development of a set of software tools that will assist the development, operation and checking of the different stages in the process of information transfer between the Producer and the Archive.



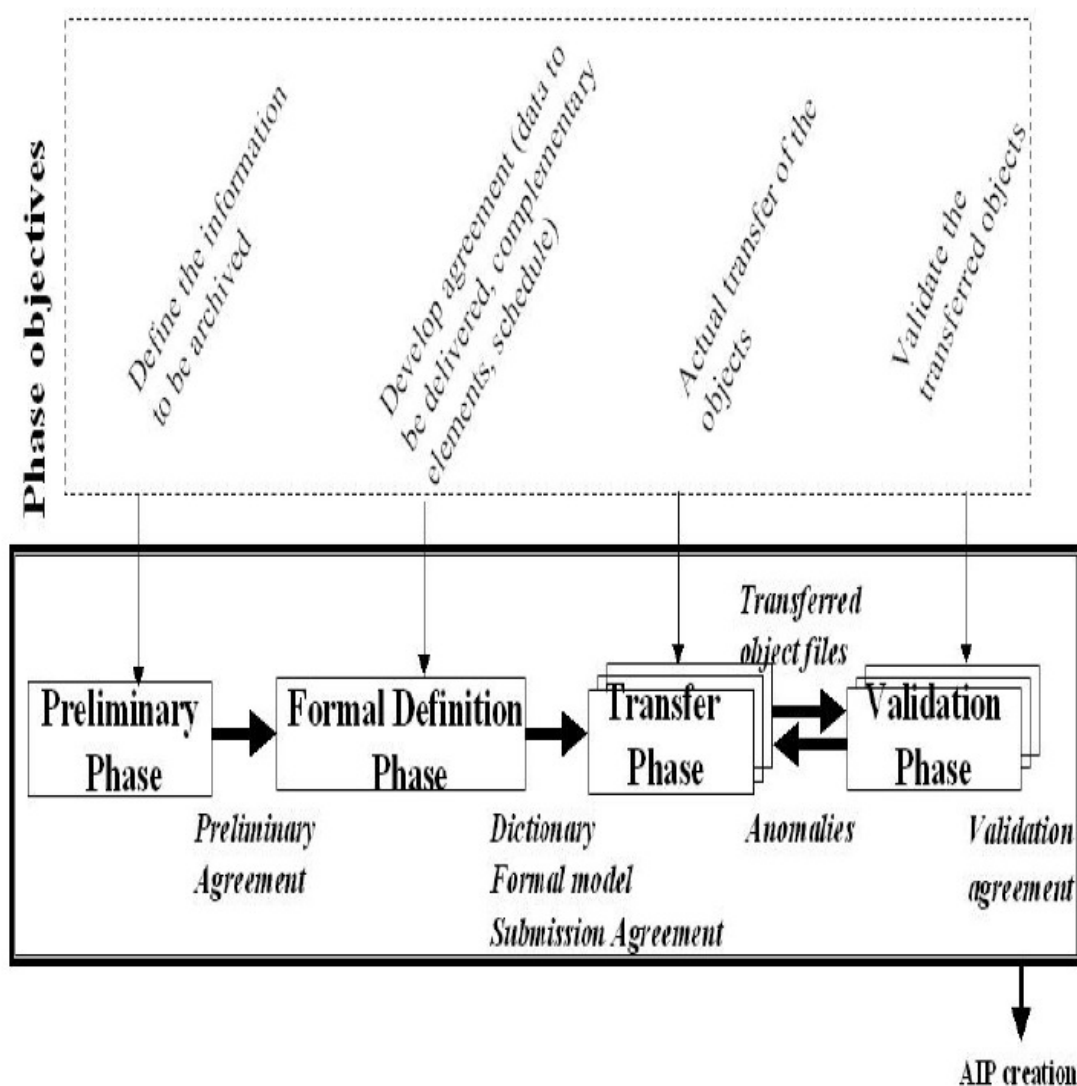


Figure 7 PAIMAS Main Phase Objectives and Outputs

- The preliminary phase leads to a summary document on the feasibility of the Producer-Archive Project and approves proceeding to the formal definition phase (or stopping the project).
- This document is the basis on which the formal definition phase is developed. The formal definition phase leads to the Submission Agreement, which summarizes all the aspects of the formal definition phase, being drawn up. This agreement refers to a Data Dictionary and a formal model. Both of these elements are needed in order to proceed with the transfer phase.
- The outputs of the transfer phase are Information Objects that are input to the validation phase. As previously mentioned, validation may be able to be started before all the Information Objects have been delivered. The transfer and validation phases are often carried out partially in parallel, as there is iteration when all of the information to be submitted is not submitted at once.





- The Archive sends the Producer its validation report for the objects received, or forms reporting the anomalies found (the Archive may also acknowledge receipt of SIPs after ingest, and only notify the Producer if there were anomaly forms or invalid data).
- There can be a significant lapse in time between the formal definition phase and the actual transfer phase. Within the Archives the transfer phase and the validation phase can take place concurrently if the actual transfer phase occurs over an extended length of time.

3.1.6 InterPARES

The InterPARES project has based its main investigation related to the function of ingesting authentic electronic records on the OAIS model: as stated in its first report published in 2003, this analysis was developed in conjunction with the San Diego SuperComputer Centre. It was aimed to support archival processes from accessioning through preservation and use, and it recognized the importance of collection-based management as expressed by OAIS reference model. The common approach concerns also the exploitation of inherent hierarchical structures within records, predictable record forms, and dependencies between them. It was also designed to be consistent, comprehensive, and independent of infrastructure as explicitly established by the OAIS reference model.

3.2 ARCHIVAL STORAGE

(A) PROJECTS AND OTHER MAJOR INITIATIVES

3.2.1 BRICKS

BRICKS offers an extensive Content and Metadata Management infrastructure based on Apache Jackrabbit reference implementation for JCR 170 Java Content Repository Application Programming Interface (API) and on Jena.

On the AIP management side this means that an AIP structure can be modeled as JCR schemas and managed at object level. The JCR 170 API allows the handling of versioning, locking, content-based search and retrieval.

JCR data structures are modeled as Nodes with associated Properties holding data, which include both elementary values and binary content. The available properties on a Node are defined by the associated mix-in types, which define as well the allowed sub-node types. Mix-in types provide a simple yet effective way of associating common properties to node, and are used also for defining behaviour for nodes, for example to support versioning, ownership and search.

Nodes are identified by a UUID which allows implementation of referential integrity, object references and OO inheritance.

The JCR repository can be queried through an XPath interface and through standardized SQL queries. Lucene provides full text search support in Jackrabbit implementation.

BRICKS has defined a set of standard types for the representation of some popular XML-based formats, such as Docbook, TEI-Lite, MPEG-21 DIDL and REL, ECHO. This proves the flexibility of the type system, which can be used to define schemas for AIPs.

Access to the resources stored within BRICKS is fully controlled by a context and role based access control infrastructure, which allows to have granular access control lists at JCR node level and at single metadata record level.





3.2.2 Chronopolis

Chronopolis⁶ is a project seeking, at the time of writing, funding in the USA, aiming to provide a model facility to enable long-term support of irreplaceable and important US data collections, ensuring that: (1) standard reference datasets remain available to provide critical science reference material; (2) collections can expand and evolve over time, as well as weather evolution in the underlying technologies; and (3) preservation “of last resort” is available for critical disciplinary and interdisciplinary digital resources at risk of being lost.

Chronopolis aims to provide tools, software, and services needed to manage data, information, and knowledge at the scales required for national digital holdings. It would function as a distributed national “data backbone” federating data and information (preservation over “space”), and would provide operational data services for maintaining key digital collections for the long-term, ranging from scientific databases to library holdings (preservation over “time”). Chronopolis plans to integrate a production system with a Research and Development (R&D) laboratory, and an administration and policy team to provide a scalable model for cyber infrastructure data management evolution and long-term preservation.

Preservation here is defined as the process of managing technology evolution and maintaining integrity by migrating to new media, new encoding formats, new information syntax, and new storage technologies as more cost effective systems become available.

The digital entities within the collections, whether they are simulation output, observational data, or derived data products, are described by a context that defines their possible uses. Context will be organized for each of the Chronopolis collections and managed as metadata attributes that include: provenance metadata (data source), administrative metadata (information about where the files are stored, file size, owner, deposition date, aggregation in a container), integrity metadata (information such as checksums, audit trails, ACs), structural metadata (information about organization of components for compound objects), and behavioural metadata (information about the encoding format).

Note that these are not the **OAIS** definitions, and in particular preservation is focussed more on structure rather than “semantics”.

3.2.3 DILIGENT

Within DILIGENT⁷, data storage is realized on local databases and on grid storage facilities. The Storage Management Service provides the basic operations for inserting, manipulating, fetching, and deleting information objects as being a generic abstraction of different concrete object types. This includes assignment of storage properties and set up of inter-object relationships. It also provides operations for associating information objects with file-base documents stored in gLite storage elements. This service is a Web Services Resource Framework (WSRF) compliant Web Service (WS).

DILIGENT repositories are trustworthy. They are resources whose participation to the infrastructure is regulated by a registration and acceptance process: the repository provider asks to add his repository to DILIGENT and the DILIGENT system administrator takes the final decision on that addition. Moreover, the access to any resource (either service or collection) is regulated by authorization and authentication covered by the Grid Secure Infrastructure (GSI) security mechanism thus preventing unauthorized accesses and unauthenticated actions.

⁶ <http://dev.dcc.ac.uk/CASPAR/pub/Main/UsefulReferences/Chronopolis--Public.pdf>

⁷ <http://www.diligentproject.org/>





3.2.4 DSpace

Massachusetts Institute of Technology (MIT) and Hewlett-Packard Labs develop the DSpace digital repository management system. It is used by many organizations around the world to store, preserve and disseminate scientific and educational content. For more details see Annex V – DSpace and CASPAR where the data storage component is discussed in more detail.

Files are represented in DSpace as bit streams. Each bit stream has an associated bit stream format, containing: a name, a Multipurpose Internet Mail Extensions (MIME) type, and a reference to the specification of the file format or, where not possible, to an application that is able to open the file. Each file format is tagged by the DSpace administrator as 'Supported', 'Known' (may be promoted to 'Supported' in the future) or 'Unsupported'.

One or more bit streams form a bundle. An item can have several named bundles. Items represent content objects, in the case of DSpace mainly scientific and educational papers. Each bundle contains different information about the item. The type of information in a bundle is told by the bundle name. The DSpace Application Layer uses names such as “ORIGINAL” (the originally submitted paper), “TEXT” (the extracted text for full-text indexing), and “CC-LICENSE” for the Creative Common license and so on. This mechanism is simplistic but very flexible, since any kind of information can be added to an item without having to alter the database design. A new kind of information simply requires a new bundle name.

In addition, each item in DSpace has also an associated Dublin Core (DC) description, which is not a bundle and is mandatory. Items are organized in collections, which are owned by designated communities.

3.2.5 Fedora

The Fedora Management Service defines an open interface for administering the repository, including creating, modifying, and deleting digital objects, or components within digital objects. The Management Service interacts with the underlying repository system to read content from and write content to the digital object and data stream storage areas. The Management Service exposes a set of operations that enable a client to view and manipulate digital objects from an abstract perspective, meaning that a client does not need to know anything about underlying storage formats, storage media, or storage management schemes for objects. Also, the underlying repository system handles the details of storing data stream content within the repository, as well as mediating connectivity for data streams that reference external content.

The Management Application Programming Interface (API-M) defines an interface for administering the repository. It includes operations necessary for clients to create and maintain digital objects and their components. API-M is implemented as a Simple Object Access Protocol (SOAP)-enabled WS.

The Management-Lite API (API-M-Lite), currently under development, is intended to provide a lightweight version of the Fedora Management Service implemented as a Representational State Transfer (REST)-based WS that can be invoked with simple Universal Resource Locator (URL) syntax. Currently, there is only one operation implemented, which is the getNextPID. A full version of this interface may be provided in a future release, depending on user demand.

3.2.6 e-Depot and DIAS

The e-Depot⁸ is the digital archiving system of the National Library of the Netherlands (KB). The e-depot was developed with IBM and is constructed using several off-the-shelf IBM products such as DB2, Tivoli Storage Manager and Content Manager. The technical core of the e-depot is an electronic Deposit system called Digital Information Archiving System (DIAS). The DIAS provides a deposit

⁸ <http://www.kb.nl/e-depot>





library solution for storing and retrieving electronic documents and multimedia files. The e-depot conforms to the **OAIS** reference standard and supports physical and logical digital preservation.

The DIAS solution allows the manual and automated ingestion of digital information (assets) into the system. Once an asset is successfully stored, it is managed for preservation and permanent access. Stored assets can be accessed either via a web-based interface (for assets having standard file types) or via a specific work environment on a reference workstation.

The Archival Storage in the DIAS is based on several products including Content Manager and Tivoli Storage Manager.

The Content Manager sub-components are the Library Server which is the metadata catalogue and the Object Server which actually holds the digital objects. The Library server uses a database to store information such as object types, indices of all stored objects, authorized system users and access control lists for each object. There can be many Object Server associated with one Library Server within the electronic deposit. The Object Server interacts with Tivoli Storage Manager which manages storage in storage pools which can be allocated either on magnetic disk, optical disk and tape.

Another important DIAS component is the Preservation Subsystem described in Erik Oltmans, Raymond J. van Diessen, Hilde van Wijngaarden, "Preservation Functionality in a Digital Archive," *jcdl*, pp. 279-286, Digital Libraries, 2004 ACM/IEEE Joint Conference on (JCDL'04), 2004. This component provides functionality for maintaining the technical metadata and the functionality needed for ensuring long term management and access to the archived publications. In the context of the **OAIS**, the Preservation Subsystem functionality is categorized as Preservation Planning. The activities of the Preservation Subsystem include monitoring changes in the technology environment, facilitating the processing to automatically migrate data from inaccessible formats and carry out permanent access strategies that will guaranty long term access to the data.

(B) OTHER TECHNOLOGIES

This section describes two new storage standards, Object-Based Storage Devices (OSD) and eXtensible Access Method (XAM). The combination of these standards enable the encapsulation of large metadata with raw data for long-lived information, thus is most suitable for Preservation Data Stores. In addition, the Storage Resource Broker (SRB) and its successor iRODS are described—these can serve as foundations for preservation projects.

3.2.7 The OSD standard

OSD enable the creation of self-managed, heterogeneous, shared and secure storage by moving low-level storage functions into the storage device itself and accessing the device through a standard object interface rather than a traditional block-based interface such as Small Computer System Interface (SCSI) or Integrated Drive Electronics (IDE). The OSD technical working group at Storage Networking Industry Association (SNIA) develops models and guidelines, requirement statements, preliminary standards definitions, reference code and prototype demonstrations for OSD storage subsystems.

The first standardization effort of an OSD specification is embodied over the SCSI protocol and is being realized as a new set of SCSI commands. Version 1 of the T10 standard was publicly reviewed and approved in late 2004; the OSD standard was published as American National Standards Institute (ANSI) INCITS 400-2004⁹. Many companies were involved and contributed to the standard. Among them are: HP, IBM, Intel, Panasas, Seagate, Veritas (now Symantec) and SUN. Presently, the standard is being extended to Version 2. Extensions include advanced functions such as snapshots, multi-object operations, collections, and error handling.

⁹ANSI INCITS 400-2004, <http://www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf>





What is an OSD object?

An OSD user-object is a container of data and attributes, indexed by a 128-bit object identifier. The object identifier is composed of a 64-bit partition ID and 64-bits object ID.

User objects are grouped into partitions where each user-object belongs to a single partition. An object can be a member of a collection of objects, all of which residing in the same partition. An object can also belong to multiple collections.

An object is created and deleted via the Create/Remove command respectively. The data of the object can be accessed via OSD Input/Output (I/O) commands (i.e., Read, Write, Append), while its attributes via the Set/Get attribute commands. I/O commands may access any number of bytes at any logical offset of the object.

User Data - An object is a sparse collection of data bytes, addressed by their logical offset in the object. That is, the object may be composed of 'holes' corresponding to offsets that were never written to with data. Semantically, such holes contain the value of 'zero'. An object has two associated parameters – its logical length and its size. The logical length is defined as the largest offset that has been written to and the size is the actual number of bytes it consumes on the physical media.

Attributes – the object's attributes are data bytes that are maintained and stored persistently with the object's data. The standard defines a basic set of attributes, some of which are programmable whereas others are not. The standard provides a mechanism to extend this basic set into user-defined attributes. All attributes are type-less.

Attributes are grouped logically into 'pages' and accessed via a pair of indices (page number, attribute number). The pages are:

- Directory page
- Information page
- Quota page
- Timestamps page
- Collections page
- Policy/Security page

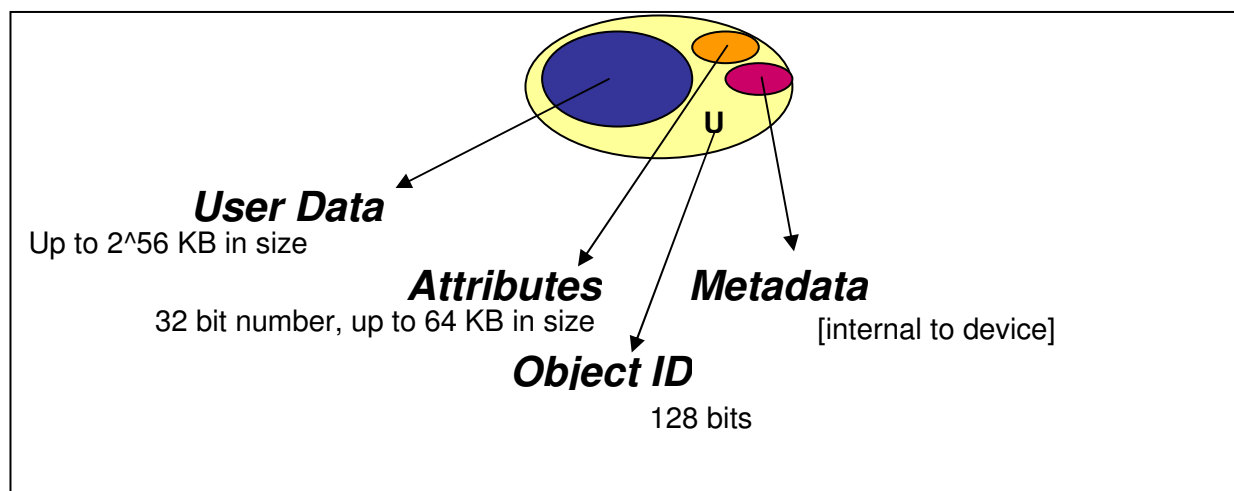


Figure 8 OSD Object

OSD Security





An important aspect of an OSD is its security model, which enforces AC on every command. The security aspect of the OSD protocol is based on a cryptographically secured capability). The Security/Policy Admin, a trusted entity, which is assumed to enforce a given policy, issues credentials.

3.2.8 The XAM standard

XAM is SNIA initiative to define a standard interface between consumers (application and management software) and providers (storage systems). It was initiated by IBM and EMC in Q4 2004 and now is supported by major vendors including Seagate, Microsoft, ByCast, Hitachi, HP, HDS and SUN.

The basic artefact in XAM is the XML search engine, XSET, record which is a data structure that is a package of multiple pieces of data and metadata that are bundled together for access under a common globally unique external name, called a XUID. The figure below is a schematic view of a XSET:

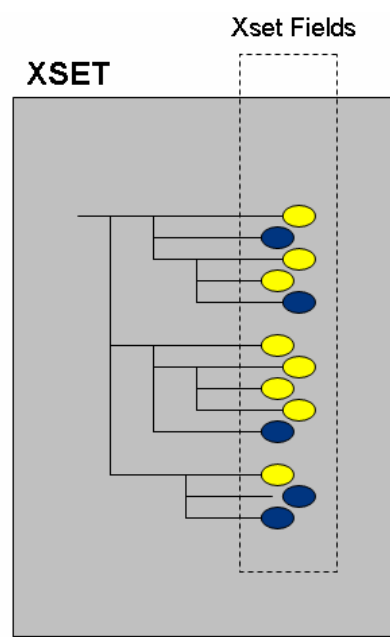


Figure 9 Schematic view of XSET

There are two types of XSET Fields:

- Properties – fields that usually include metadata and thus will be indexed and used in queries. Their type is a “simple” type and is one of Boolean, Int64, Float64, String256, DateTime, XUID. The property type is checked and enforced by the storage system. The API to these fields is via get/set methods.
- Unstructured Streams (u-streams) – fields that include unbounded byte streams. Their type is a valid MIME-type but the type is not checked or enforced by the storage system. The API to these fields is based on Posix I/O methods i.e., open, read, write, and close.

Future versions of XSET may introduce a new field called Structured Streams (s-streams). These are similar to u-streams in that they include unbounded byte streams and are manipulated by Posix like methods, but their type is expected to be a XML schema and validation with that XML Schema is checked and enforced by the storage system.





Each field has a fixed set of attributes for its content and behaviour that are manipulated via get/set methods. The attributes are:

- Type – the type of the value of the field namely simple type for properties, MIME-type for u-streams
- Value – the actual value (content) of the field
- Fixed - a Boolean value indicating if the field is immutable within the XSET; that is, if the field can be modified without the automatic creation of a new XSET with a different XUID
- Read-only - a Boolean value indicating if the field can be modified by an application.
- Length – the actual size of the field value in bytes. This attribute value is provided by the XAM system and is read only for the application.

The XUID, which uniquely identifies a XSET, is a variable-length byte string, up to 80 bytes in length. It includes the Simple Network Management Protocol (SNMP) enterprise number of the storage vendor according to RFC 2578 (i.e., IBM=2, NetApp=789, EMC=1139) and an opaque id unique within the storage vendor. Where applicable, XUID textual representation should be base64-encoded, as described in RFC 2045.

3.2.9 Storage Resource Broker (SRB)

The Storage Resource Broker (SRB)¹⁰ is a data grid technology, developed and owned by the San Diego Supercomputing Center (SDSC). It manages distributed data, enabling the creation of data grids that focus on the sharing of data, and was recently extended to persistent archives that focus on the preservation of data. Data grid technology provides the fundamental management mechanisms for distributed data in a scaleable manner. It targets environments that deal with combinations of distributed data sources, distributed data curators, and distributed users. It has the ability to assemble a shared set of data whose records reside in multiple types of storage systems, at multiple institutions and at multiple geographical sites. It includes support for managing data on remote storage systems, a uniform name space for referencing the data, a catalog for managing information about the data, and mechanisms for interfacing to the preferred access method. The SRB is a middleware software - it builds on top of standard file systems, commercial archives and storage systems.

Over the past years, SRB has been used as a foundation technology for providing a persistent archive in the context of preservation projects. Many of these projects were commissioned by NARA, the National Archives and Records Administration, and supported by the Library of Congress and NSF. As a persistent archive, it managed the retention of the digital record as well as the context that describes the origin, relevance and authenticity of the record.

SRB focuses more on the bit preservation aspect of the problem. It handles media failures, data mirroring and distribution of data. It stores the data records as files on the storage system and assumes that the data object is packaged into an AIP. The AIP is written to the storage repository but a separate database is used to store the metadata related to the electronic record.

¹⁰ http://www.sdsc.edu/srb/index.php/Main_Page



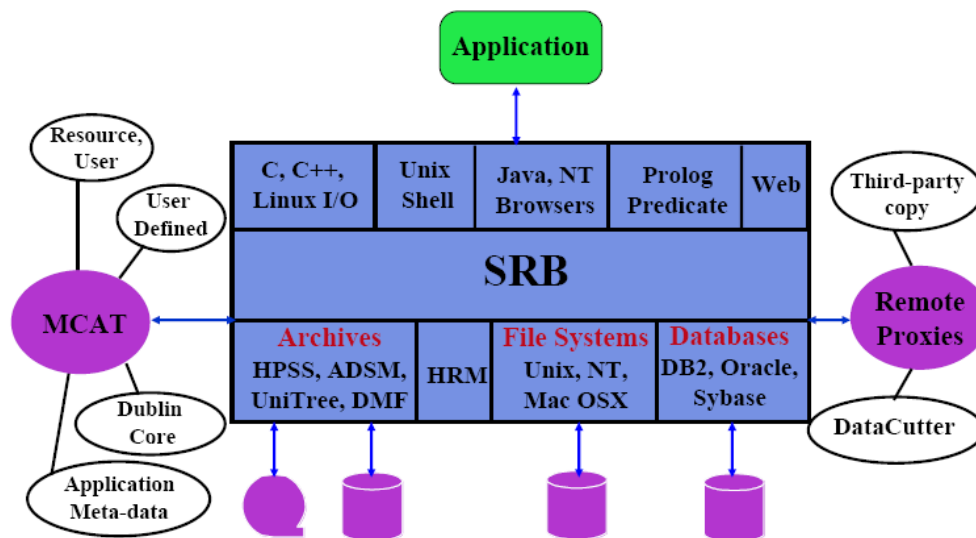


Figure 10 Architecture of the Storage Resource Broker

The basic architecture of SRB consists of an SRB server and a Metadata Catalog (MCAT) server. The SRB server exposes various file-like APIs to the application and interacts with the storage system. The MCAT server handles the information stored in the SRB database. The picture above is taken from SRB documents.

3.2.10 iRODS

iRODS¹¹ which stands for *i* Rule Oriented Data Systems, is a new data grid software being developed by the San Diego Supercomputing Center (SDSC) and other contributors. It builds on the expertise gained from the application of the SRB software to persistent data archives. It is viewed by its developers as the continued evolution of the basic control mechanisms required to manage distributed environments. The new software has been released in December of 2006 and not much experience has been gained with it so far.

iRODS explores the idea of using *constraint and rule-based systems* for the management of distributed data based on distributed grid models such as SRB. It aims to characterize management policies, and to automate the application of these policies. The management policies are mapped onto rules that are applied on the execution of all data management operations.

There are a few examples for the application of rules to persistent archives of data.

- The automation of the evaluation of trusted digital repository assessment criteria¹². The rules can be used to explicitly characterize the management policies as well as to provide a mechanism to track the results of applying the management policies.
- Define a mapping of the set of capabilities required by the NARA Electronic Records Archives to rules that control execution of fundamental services¹³. It is reported that all 854 listed capabilities were defined by only 174 rules applying 212 metadata attributes for the implementation of the ERA capabilities.
- Define Infrastructure management of shared collections within data grids.

¹¹ http://irods.sdsc.edu/index.php/Main_Page

¹² Audit Checklist for Certifying Digital Repositories, http://www.rlg.org/en/page.php?Page_ID=20769

¹³ AERA capability requirements as of 8/11/2006, <http://www.archives.gov/era/pdf/requirements-amend0001.pdf>





iRODS inherits the basic SRB mechanisms for data virtualization, resource and storage virtualization, trust virtualization and the various standard APIs. In addition, it allows workflow and management policies to be executed. At the heart of iRODS is a *Rule system*, with a *Rule Engine* that interprets rules on how the system should react to various conditions and requests. This rule system can be used to express management policies. iRODS provides tools to support dynamic construction of rules as needed by the client. Rules can be run in many different modes, for example immediate execution and delayed execution.

3.2.11 InterPARES

At a high level, it may be said that the specific component of InterPARES identified as ‘Preserve Electronic Records’ model is a specification of an OAIS for the specific classes of information objects comprising electronic records and archival aggregates of such records. By the way the project is narrower than the OAIS model in the sense that the InterPARES preservation model does not include all activities related to making records available—only those that are inextricable from the preservation function: for instance it does not include order agreements as described in the OAIS model or any ‘value-added’ dissemination or access services. Similarly, the preservation model does not include processes, which inform potential users what records are being preserved or what conditions govern access to the records.

As clearly stated in the final report in InterPARES 1, while the ‘Preserve Electronic Records’ model is narrower than the OAIS model, the InterPARES model has substantially more depth on the topic of preservation in general and, obviously, the preservation of authentic electronic records in particular. The Preservation Task Force communicated its work to the committee responsible for the OAIS standard and worked with that committee to enhance the standard in light of the project findings.

3.3 DATA MANAGEMENT

(A) PROJECTS AND OTHER MAJOR INITIATIVES

3.3.1 BRICKS

The BRICKS Data management services include the following “bricks”, in the project jargon:

Collection manager brick: this service provides the main interface for all content and metadata-related operations. Collections (sets of sub-collections, items and references to items) are handled by the Collection Manager and used to organize DLObject (digital library objects) in a hierarchical manner.

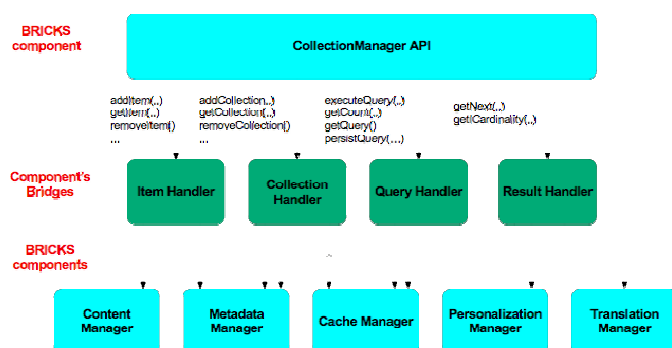


Figure 11 BRICKS Collection Manager Architecture

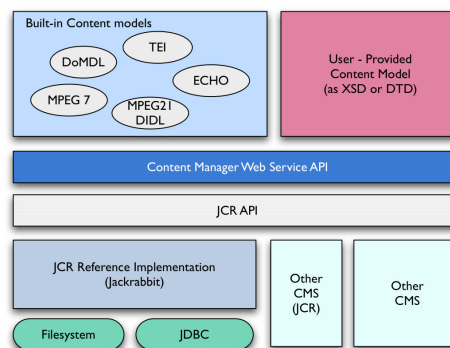


Figure 12 BRICKS Content Manager architecture





Content manager brick: this service is used as a backend for storing and managing the content part of BRICKS DObjects, allowing fine grained editing through the Java Content Repository API (JSR 170). This service can handle arbitrary complex content models through the JCR type system. Content is imported in the BRICKS content manager by copying bit streams into the filesystem based storage or by referencing source URLs. It should be noted that a user who accesses a document will not notice if the Content Manager stores the actual content or just a reference to it.

Metadata Manager brick: the service provides a repository for all descriptive metadata about BRICKS Collections and DObjects. The descriptions use the Resource Description Framework (RDF), not exposing directly the RDF Graph (which can be accessed through an export function), but intelligent data structures called Records, organized by schema namespaces and containing the information objects. Metadata schemas are managed through ontologies represented as OWL, a widely adopted standard developed by W3C, and shared through the **Ontology Manager brick**, that provides ontology deployment and mapping services.

Each brick provides its functionality through a SOAP Web Service interface. All operations are controlled by Access Control Policies, embedded from the Web Service provider implementation.

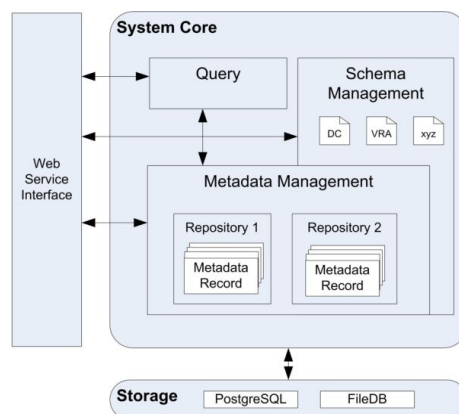


Figure 13 BRICKS Metadata Manager Architecture

3.3.2 DILIGENT

Data management consists of the following services: Storage Management, Content Management, Metadata Management, Metadata Broker, Content Security and Annotation Management.

Content and Metadata Management Services are devoted to the data and metadata management. These services (1) provide access to both the information objects and their metadata, (2) encrypt and decrypt them, and (3) manage their annotations. To this end these services are logically grouped into:

- Content Management
- Metadata Management
- Content Security
- Annotation Management

Content Management Services provide the high-level operations, which are mapped onto generic Storage Management operations. These operations are: (1) basic documents storage, access and update, (2) collection management, (3) fundamental metadata (storage properties) association, (4) archive import and (5) notification management. These services are WSRF compliant WS.

The Metadata Management Service is composed by:

- The metadata catalogue service (WSRF service), responsible for managing the incoming requests for changing, accessing, manipulating metadata. It (1) relates metadata with information objects,





(2) removes associations, (3) returns metadata for a given object and (4) updates associated metadata.

- The metadata broker service (WSRF service), it provides the functionality for importing and transforming new metadata and therefore managing the involved transformations. To be able to perform this task, the service offers management capabilities to store, associate, find and delete transformation rules, which can be applied to transform metadata from a source to a target schema. Additionally, transformation programs must be managed and maintained. They can be used to perform a series of transformation processes.
- The metadata management administrative user interface. The metadata management is working in background and called by other services, thus it has been equipped with a portlet providing administration and configuration functionalities.

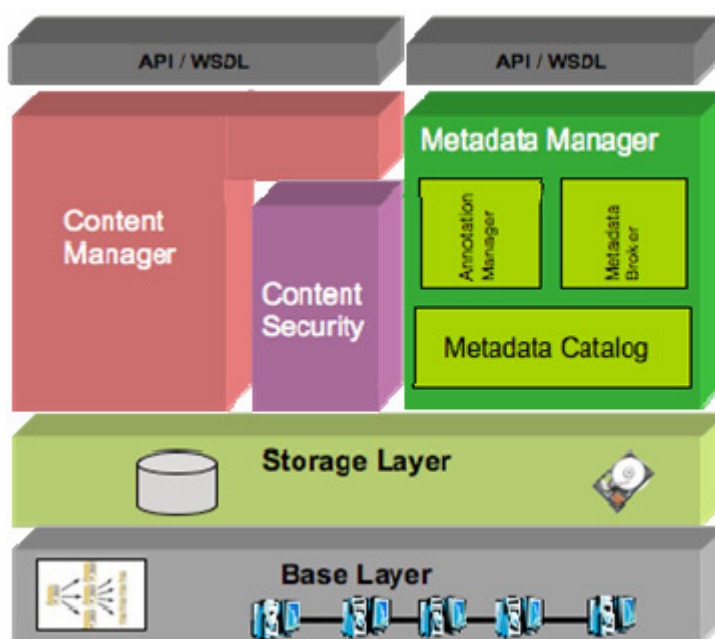


Figure 14 DILIGENT Content & Metadata Management: Layered Architecture

3.3.3 Fedora

The Data Management Database service corresponds closely to the **OAIS** data management functional entity. Fedora 2.1 comes with two mechanisms for discovering objects in the repository.

The first mechanism is called “Basic Search.” It is a simple text-based search of the basic properties of the Fedora Object eXtensible Markup Language (FOXML) that wrap Fedora objects and the DC metadata. The basic search is simplistic.

The second mechanism for discovering Fedora objects is the RDF triple store. Every Fedora object can have a distinguished metadata stream, which contains RDF statements. These RDF statements are maintained in a triple store, which supports arbitrary Interactive Tucana Query Language (iTQL) queries. It is additionally possible to maintain external search indexes. Maintaining such an external index requires either (1) using an external application which manages both ingest and the index or (2) keeping the index in sync with the repository manually. An Alerting Service could be utilized and would allow one to automatically maintain such an external index.





(B) OTHER TECHNOLOGIES

3.3.4 iRODS

The discussion of iRODS under archival storage includes some examples of rules used to express policies for data management.

3.4 PRESERVATION PLANNING

(A) PROJECTS AND OTHER MAJOR INITIATIVES

3.4.1 BRICKS

BRICKS is not intended for directly addressing long-term preservation of the Digital Library contents, but provides a number of features that can allow the integration of dedicated solutions for digital preservation. BRICKS allows to manage an arbitrary number of schemas for descriptive information, with a mandatory subset of DC for titles, descriptions, author information, copyright, etc. The MIME-Type field provides the representation information for bit streams.

BRICKS provides identifiers expressed as unique persistent URNs, adopting some restrictions in order to include information about the BRICKS node managing the referenced DLO and its location in the collection hierarchy for that node, by using the following syntax:

Access control policies are bound to these persistent identifiers and represented through XACML (eXtensible Access Control Markup Language), a widely accepted and implemented standard.

3.4.2 DILIGENT

DILIGENT's approach to 'digital preservation' is to rely on (1) the huge storage capability provided by GSI, which allows multiple copies to be stored, and on (2) the computing capability provided by the grid to run ad hoc user defined jobs. This will allow the dynamic generation of further manifestation formats of the objects to be stored. As DILIGENT provides an infrastructure, its goal is to support the preservation policies to be implemented in the application scenario.

3.4.3 DSpace

3.4.3.1 Preservation in DSpace

DSpace does not address issues related to the long-term preservation of the content it stores. Built-in preservation metadata map to the DC scheme (i.e., various contributors, provenance information, creation date, etc.) The current DSpace version is a work in progress and many steps forward are foreseen for DSpace version 2.0.

That said some very fundamental problems are solved. The most basic is the persistent identification system, ensuring stable and resolvable identifiers for stored items. DSpace uses Corporation for National Research Initiatives (CNRI) Handle System, which appears to be the most Uniform Resource Name (URN)-compliant persistent identification system currently available.

DSpace also offers a history system where changes to items are recorded in RDF using the Harmony-ABC ontology (which is however now obsolete and unmaintained). Using this system, items can be (1) deleted completely or (2) withdrawn without deleting any information from the database.

Bit stream representation information is provided by the MIME type and by references to specification documents and/or software applications. This is quite basic but complete: you can reference a format specification or a software tool and both could be stored as items on their turn.

Checksums for bit integrity check are also stored in the database. The bit streams themselves can be stored on file system or using Storage Resource Broker (SRB), a very complete distributed storage





management system. This allows data storage on heterogeneous supports, provides a separation between the physical and logical storage location, as well as replication functionalities.

3.4.4 Fedora

Fedora has two capabilities that are relevant to preservation functions.

- Security architecture and policy enforcement: Fedora provides a pluggable authentication module using Tomcat's standard approach to authentication, as well as a new AC module that enforces policies written in eXtensible Access Control Markup Language (XACML). Fedora has two plug-in modules for authentication: (1) a standard module that authenticates using a file of user identity and role information (i.e., tomcat users.xml) and (2) a Lightweight Directory Access Protocol (LDAP) module to obtain user attributes from an LDAP directory. Fedora also provides a XACML-based policy enforcement module for authorization purposes.
- Resource Records. The Ingest and Maintain Guides rely on resource records including format information, record type information, submission agreements, producer metadata, retention schedules, knowledge base metadata, and the history of the repository itself. These supporting records can also be represented in a straightforward manner as Fedora objects with content models describing the kinds of information and abilities expected from such objects. Fedora also allows data stream versioning, which is a very important capability for resource records whose content may change periodically. Encoded Archival Context may be adapted as a standard way for encoding producer records, but an alternate method to encode information for the remaining supporting record types remain to be developed.

3.4.5 MUSTICA

3.4.5.1 Benefits and limits of MUSTICA in musical work preservation

First, MUSTICA gives the composer and/or his assistants an opportunity to record and preserve, in standardized (XML Schema validated) syntax, information that would typically be retained only in their minds. This includes a precise description of the roles and relations of all the performance files as well as the equipment and software needed. This is an advancement compared to current practices in contemporary music, particularly if a composition includes live electronics, performed under the supervision of the composer and his assistants. It is very hard to re-perform it without their help even if technologies are still available.

The next issue, preservation despite software/hardware changes and alterations in the knowledge base of the designated community, is addressed but not resolved. MUSTICA documentation contains guidelines for the format(s) of common file types (audio, images, video, text documents etc.). The XML Schema also contains a simple formatted text definition with the possibility to include links to the digital assets, which is a kind of virtualization for performance instructions.

Technical schemes are usually just images or Portable Document Format (PDF) documents that use graphical conventions not always understood by other applications. This can even happen for score files. A composition often contains exotic file formats for specific software or refers to specific hardware, which may become unavailable in time. This has already occurred for many compositions of the 20th century. A step forward in modelling standard document types could reduce but not solve the problem, as composers can use any kind of software tools and not all can be modelled exhaustively. Migration is a key point for this reason.

3.4.5.2 Migration of musical works in MUSTICA

MUSTICA assumes that migrations are done manually. This is a weak but realistic assumption that provides a way to record various versions of a work, and does its best to make available all information that can help in porting a work to other technologies (recordings of previous executions are crucial in this context).





The migration concept is an important part of MUSTICA, but it looks quite unexpected from an **OAIS** point of view. The migration of a musical work towards newer (and possibly more standardized) technologies is a very sensitive point, as transcriptions in traditional music. It can be very difficult from a technical point of view, since you might not know exactly what a given software/hardware does (or did), and even very slight changes can compromise the authenticity of the work. Furthermore (and also in consequence of this) each migration will cause decisions to be taken that are not just technical but involve relevant artistic aspects.

3.4.6 Preservation planning for virtual arts

With regard to preservation of art works, a number of preservation strategies have been studied under the following projects:

- Archiving the Avant-Garde¹⁴ and its two case studies / related projects: “Renewing the Erl King”¹⁵ and “Preserving the Rhizome ArtBase”¹⁶,
- The Variable Media Network project¹⁷, and
- Digital Video Preservation Reformatting Project¹⁸.

3.4.6.1 Archiving the Avant-Garde

The Archiving the Avant-Garde project is an ongoing project, which is focused on preservation of variable media art, such as performance, installation, conceptual, and digital art. It particularly looks at practices and standards of metadata that directly support the preservation process. It investigates the use of the hardware (i.e., chipset) emulation method for preservation of avant-garde artworks. In this project, hardware emulation is considered as the key factor for preservation of art work based on the argument that “hardware is well documented on a technical level, and that this approach gives us the most leverage for the investment” and that with “just a few hardware emulators, we could run dozens of operating systems, thousands of applications, and millions of documents”. “Renewing the Erl King” and “Preserving the Rhizome ArtBase”, two related projects, aim at applying the hardware emulation strategy in practical applications. However, no hardware emulator has been built. The “Renewing the Erl King” project planned to develop a hardware emulator, but, due to technical difficulty, original source codes were re-interpreted to re-produce the artwork instead. The emulation strategy was intended for use in conjunction with storage migration in “Preserving the Rhizome ArtBase” project. Experiments and testing on emulation software has not occurred to date, due to the expense and time involved.

3.4.6.2 The Digital Video Preservation Reformatting Project

The Digital Video Preservation Reformatting Project by the Dance Heritage Coalition focussed on the preservation of dance videotapes. Its aim is to preserve the media that captured the works versus the works themselves. To this end the project also conducted research into candidate formats for the digital video migration of content in older video formats.

¹⁴ Archiving the Avant-Garde, "Archiving the Avant-Garde: Documenting and Preserving Variable Media Art," 2006, http://www.bampfa.berkeley.edu/about_bampfa/avant_garde.html

¹⁵ J. Rothenberg, "Renewing the Erl King," 2006, http://www.bampfa.berkeley.edu/about_bampfa/ErlKingReport.pdf

¹⁶ R. Rinehart, "Preserving the Rhizome ArtBase," 2002, <http://rhizome.org/artbase/report.htm>

¹⁷ "Variable Media Network," <http://variablemedia.net/e/welcome.html>

¹⁸ Dance Heritage Coalition, "The DHC Digital Video Preservation Reformatting Project," <http://www.danceheritage.org/preservation/digital.html>





3.4.6.3 The Variable Media Network

The Variable Media Network proposed an unconventional new preservation strategy for preservation of world-renowned collection of conceptual, minimalist and video art. It encourages artists to define their artworks independently from medium so that the artworks can be recreated at another time, when the medium becomes obsolete. The Variable Media Network also aims at developing the tools, methods and standards needed to implement this strategy. In this project, a number of case studies have been conducted to compare different preservation approaches, i.e., storage, migration, emulation and re-interpretation. Storing all materials related to an artwork is only the last resort option. A known disadvantage of migration method is its inability to guarantee the conversion of data during migration processes. However, a case study in this project showed that in some situations, both emulation and migration methods were equally preferable. A case study on emulation of digital artworks, in this project, showed that emulation is never easy. Although emulation had been known as a method that could create the original look and feel of the original work¹⁹, results of the case study showed that it was not easily achievable, due to many differences between the original hardware platforms and their emulated counterparts, such as Central Processing Unit (CPU) speeds, look and feel of the new hardware platforms²⁰. Re-interpretation seemed to be a preferred approach, where the artworks could be encoded using scores and scores of music²¹. These scores would form a basis for reconstruction of the works, like replaying a piece of Mozart's music from written music scores. Re-interpretation may not produce exact replicas of the original works. However, it does give liberty to the people who will reconstruct the works and the main theme of the works can still be preserved.

3.4.6.4 Summary of methods for preservation planning for virtual arts

Research into the preservation of artworks has investigated multiple preservation methods. No absolute answer on which is best has been found. Rather, depending on the contents and requirements of preservation, one method may be more preferable than another. Advantages and disadvantages of each method are summarised below:

- Emulation strategy: preserve the data in its original format together with the application that uses the data. When the platforms on which the application can run become obsolete, emulation engines will be used. With this method, either the preserved applications have to adapt to the emulation engine or vice versa. The emulation engine itself is also dependent on the hardware platforms on which the application ran. If the hardware is changed, the emulation engines need to be recoded. In this case, the preservation is actually about preserving the applications and the emulation engines. An advantage of this approach is that the look and feel of original applications can potentially be preserved. A disadvantage is the lack of interoperability between the preserved data, applications and associated emulation engines. If one of these components is missing, the reconstruction may not be possible.
- Migration strategy: preserve the data in its original format and convert the data into usable format when the original format becomes obsolete. The preservation is about preserving the data. New application can be developed to interpret the data when necessary using data format

¹⁹ J. Rothenberg, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation," Council on Library and Information Resources 1999, <http://www.clir.org/pubs/reports/rothenberg/contents.html>.

²⁰ "Seeing Double: Emulation in Theory and Practice," Solomon R. Guggenheim Museum, 2004, <http://www.variablemedia.net/e/seeingdouble/index.html>.

²¹ R. Rinehart, "A System of Formal Notation for Scoring Works of Digital and Variable Media Art," presented at Annual Meeting of the American Institute for Conservation of Historic and Artistic Works, Portland, Oregon, 2004.





documentations. However, conversion is not always complete. Even with a specific format documented by a reliable source, errors can still exist. Consequently, the reconstructed look and feel may not be identical to the original version.

- Use [of] standard formats: at data level. This preservation is about the preservation of images, audio and video contents. Even the best data format may be changed over time. However, in short-term preservation, standards can help to make applications, which conform to a common standard interoperable.
- Re-Interpretation: is an approach to preservation of variable media artworks proposed in the Variable Media Network and allows the re-creator flexible reconstruction of the preserved contents. However, this method will not exactly reproduce the original works.

In the case of the Performing Arts Test-bed for **CASPAR**, the interests are in two areas. Firstly, it is the process in which a preserved content is put in storage for archive and retrieved at a later time for reconstruction. Secondly, it is how the preserved content (i.e., a performance, including its environment setting, the procedure, related hardware and software) can be reconstructed as precise as possible the same as its original version. This is what we are particularly interested in.

Further information is available in Annex III – Preservation of Interactive Multimedia Performances and 3D Motion Data Representation.

3.4.7 SAFE

The volume of Earth Observation (EO) data archived by European Space Agency (ESA) exceeds 1.5 PB and is constantly increasing with the on-going operations of European Remote Sensing Satellite (ERS), Environmental Satellite (ENVISAT) and Third Party Missions. It is predicted that this volume will exceed by more than 2 PB within a few years. New datasets from the future Earth Explorer and Global Monitoring for Environment and Security (GMES) missions will continue to increase the volume in the years to come. All these datasets are archived for the long-term in several centres around Europe and are used to generate systematic and on demand end user products. The ESA's mandate is to preserve datasets under its ownership for ten years after the end of the mission's active life. On the other hand, it is recognized that the data to be preserved long-term requires special attention. This is reflected in costly operations for their exploitation and maintenance. The excessive proliferation of diverse and heterogeneous data formats is the result of:

- The lack of agreed standards in the EO community, as the formats tend to be specific for the sensor(s) carried on board each mission.
- Legacy data from old ground segments architectures, which tend not to reuse elements previously developed.
- Until recently the information technologies and standards used to describe and package the data were immature, preventing the creation of a unique format able to satisfy the requirements for the long-term data preservation and also their handling in the processing centres.

The ESA's EO Department has recognized the need to standardize and harmonise its ground segments architectures to gain economies of scale during development, operations and maintenance, including the need to standardize the formats in which the datasets are preserved. By achieving this goal, the way would be also paved for simplifying the exchange and interoperability of data between ESA and external operators. In early 2004 ESA launched a project called Historical Archives Rationalization and Management (HARM), which aimed at converting its historical datasets into a new modern format, based on the latest technologies and standards in order to ensure the long-term preservation of its holdings.





SAFE is designed to act as a common format for archiving and conveying data within ESA EO archiving facilities²². As such, SAFE benefits from the experience gathered while developing standards related to data formats. SAFE intends to resolve the major challenges coming from the packaging and the long-term preservation of EO data. Special attention has been made to ensure that SAFE conforms to the ISO 14721:2003 **OAIS** reference model and related standards like the emerging Consultative Committee for Space Data Systems (CCSDS) / ISO XFDU (XML Formatted Data Units) packaging format. Although the primary goal of SAFE, in the framework of the HARM project, is to handle EO data with processing levels close to what is usually called “level 0”, no limitation exists regarding the packaging of higher level products as well as other technical and scientific information. Actually, experience has demonstrated that packaging and archiving higher processing levels or auxiliary data in a common format may be effective in many situations. SAFE undergoes this concept by offering a single framework for packaging a large variety of information.

3.4.8 InterPARES

The “Chain of preservation” as approved at the conclusion of InterPARES 2 (2006) is based on the idea that the planning of preservation cannot be confined to the end of the chain of the digital resources. This statement is strongly related to the fact that the authenticity of electronic resources is threatened whenever they are transmitted across space i.e. when sent to an addressee or between systems or applications, or time i.e. either when they are in storage, or when the hardware or software used to store, process, or communicate them is updated or replaced. Therefore, the preserver’s *inference* of the authenticity of electronic resources, as relevant component of the preservation function, as to be planned as soon as possible and can be supported by evidence, provided in association with the resources through its early and continuing *documentation*, by tracing the *history* of its various migration and treatments which have occurred over time. Evidence is also needed to prove that they have been maintained using technologies and administrative procedures that either guarantee their continuing *identity* and *integrity* or at least minimize risks of change from the time the resources were first set aside, to the point at which they are subsequently accessed. According to the InterPARES (but not in contradiction with the OAIS approach) authenticity is never limited to the resource itself, but is extended to the information/document/record system, and thus to the concept of reliability. Authenticity is also concerned with control over the information/document/record creation process and custody. The verification of the authenticity of a resource is related to the reliability of the system/resource, and this reliability should prove that it is fully documented with reference both to the creation process and to the chain of preservation. The planning of the preservation has to be defined as a chain of activities and a sum of information collected in the course of the resources management, as soon and complete and accurate as possible and as automatically as possible. In general the resource/record has to be:

“Carefully managed throughout its entire existence to ensure that it is accessible and readable over time with its form, content, and relationships intact to the extent necessary for its continuing trustworthiness as records” (from the InterPARES report “Managing the chain of preservation model - MCP”)

It is relevant to stress that all the phases and steps are interconnected to make preservation feasible and successful. It is also recognised that authenticity cannot be viewed as a stand alone characteristic of the resources, but that it is part of an overall process and an overarching concept, defined in the more recent outcome of the InterPARES research as trustworthiness of the digital resources.

3.5 ACCESS

²² <http://earth.esa.int/SAFE/>





(A) PROJECTS AND OTHER MAJOR INITIATIVES

3.5.1 BRICKS

BRICKS offers an extensive, ontology driven search & browse infrastructure. Any digital object managed by BRICKS can be retrieved based on: - (1) its Collection identifier, (2) a simple-text search, or (3) on complex metadata schema-based searches involving one or more sets. The ontology support allows providing support for an arbitrary set of metadata. The search infrastructure also supports multilingual searches and search result personalization based on user profiles.

The BRICKS Query Language provides the following types of queries: a simple search, which is a full-text search on metadata records, and can be multi-lingual, an advanced search, a boolean combination of simple (attribute operator value) conditions, on fields of a metadata schema, an ontology search, specifying conditions on the membership of the sought DL objects to a concept expressed in a certain ontology, structure-based search, and finally mixed search, allowing the combination of two or more of the above query types in a single query.

Personalization adds the possibility to perform, at query evaluation, the adaptation of query results to reflect the preferences expressed by the user in its profile.

3.5.2 DILIGENT

Following the user's request to retrieve or index information, DILIGENT provides data access services, which orchestrate the operations provided by other services. Data Access services consist of:

- Search;
- Content Source Description and Selection;
- Content Indexing;
- Personalization;
- Feature Extraction.

3.5.3 Fedora

The Fedora Access service defines an open interface for accessing digital objects. The access operations include methods to provide reflection on a digital object (i.e., to discover the kinds of dissemination available to the object), and to request dissemination. The major function of the Fedora Access service is to fulfil a client's request for dissemination. To support disseminations, the underlying repository system must evaluate the behaviour associations specified in a digital object, and determine how to dispatch a service request to a supporting service associated with the digital object. The supporting service may be internal to the repository system, or it may be an external WS that the repository must call upon. The underlying repository system facilitates all external service bindings on behalf of the client, simply returning a dissemination result via the Access service layer.

Fedora provides an Access API (API-A), which defines an interface for accessing digital objects stored in the repository. It includes operations necessary for clients to perform disseminations on objects in the repository and to discover information about an object using object reflection. API-A is implemented as a SOAP-enabled WS.

Further an Access-Lite API (API-A-Lite) defines a lightweight version of the Fedora Access Service that is implemented as a REST-based WS that can be invoked with a simple URL syntax.

(B) OTHER TECHNOLOGIES

In addition to the work on access interfaces in particular projects, there have been a number of efforts to provide a common access interface to digital repositories which allow the federation of access across archives.





3.5.4 OpenURL

OpenURLs are URLs with a set of standard search parameters, such as author and title. It is an international / ANSI standard, developed mainly for giving a standard solution to content providers in publishing more long living references to their content objects than can be achieved using common HTTP addresses.

OpenURL is, at least, a standardization of a common paradigm, which consists in accessing a resource by calling a WS with appropriate search parameters. Supporting OpenURL for data access increases the accessibility to the data themselves and can be a basis for interoperability with other systems. OpenURL is widely used, particularly by libraries and educational institutes. Several OpenURL resolvers, called linking servers, are available on the market.

3.5.5 OAI-PMH

The Open Archive Initiative (OAI), funded by the Coalition for Networked Information (CNI) and the Digital Library Federation (DLF), has developed the Protocol for Metadata Harvesting (OAI-PMH), whose definition has been stable since 2002. The paradigm is the opposite, although complementary, to OpenURL, since it provides a standard way for a data repository to expose metadata to third party indexing engines. These can collect information and provide access services about data located in several repositories. This is an interesting paradigm since it clearly separates storage and accessibility of data, allowing different subjects to work out these tasks separately, and even independently.

OAI-PMH is based on XML and allows repositories to expose several different metadata formats provided a XML Schema specifies them. The DC metadata scheme is mandatory for every repository. Standard profiles can then be defined to allow a richer metadata for particular domains or object types.

3.5.6 Semantic web standards

The RDF and the OWL layered on top of it are the result of an effort to bring the benefits of formal logic to the World Wide Web, de-coupling data from the applications and file formats it was originally created for. (For a detailed discussion refer to Annex XI – RDF and OWL.) RDF and RDF Schema provide a model for describing objects (both real-world and electronic) in terms of classes and properties drawn from a distributed set of vocabularies (often termed ‘ontologies’). RDF (and RDFS/OWL) allow objects to be uniquely identified with URIs Uniform Resource Identifiers (URIs), as well as providing a formally-grounded way to identify things via their descriptions. OWL allows more complex ontologies (typically restricted to Description Logic) to be built to provide a richer description of those resources and properties.

The main problem of RDF/OWL is its weak ability to validate data for incompleteness and in error detection. This is mainly due to the open world assumption and to its non-adopting of the unique name assumption. To be suitable for use in AIPs, additional forms of data integrity checking are needed which go beyond those offered by RDFS/OWL tools. A method based on SPARQL or RQL patterns could be suitable, but the current state of such technology is one of pre-standards exploration.

The main strengths of RDF/OWL include (1) a common access language, which is widely understood, (2) its ability to combine and merge data from different sources, using standard Query Languages (QLs) such as SPARQL Protocol and RDF Query Language (SPARQL) and the emerging Rule Language. Thus is suitable for the Access function of the AIS, because of its ability to model the knowledge base of the designated community. If the designated community knowledge base is modelled in this fashion, AIPs can be accessed, shared, disseminated and combined using these ontology elements. For detailed information please refer to Annex X - Semantic Web Knowledge Management Components.





3.5.7 XML/XSLT driven dissemination

The MUSTICA project, as well as many others, uses eXtensible Stylesheet Language Transformations (XSLT) transformations to publish the content object descriptions in different formats (currently HyperText Markup Language (HTML) or PDF). XSLT is a flexible language that can transform XML documents into new XML documents containing a subset of the same information encoded in a different syntax (for example from the MUSTICA XML representation to eXtensible HyperText Markup Language (XHTML)). The advantage of this technique is that data and presentation are clearly separated and that new dissemination formats can be supported simply adding new XSLT stylesheets.

The eXtensible Markup Language Query (XQuery) is a more recent development from the World Wide Web Consortium (W3C). This language has many features of similarity to XSLT, but it is tailored to the native querying of data from XML databases.





4





5 OAIS INFORMATION MODEL

5.1 INFORMATION MODELS

(A) PROJECTS AND OTHER MAJOR INITIATIVES

5.1.1 Cedars

The Cedars (CURL Exemplars in Digital ARchives) Project ran from April 1998 until March 2002. Funded by JISC (the Joint Information Systems Committee of the UK higher education funding councils), as part of its Electronic Libraries (eLib) Programme, Cedars was the only project in the programme to focus on digital preservation. A collaboration between three CURL institutions (the universities of Leeds, Oxford, and Cambridge), it also involved a number of other key stakeholders such as non-CURL libraries, the National Preservation Office (NPO), the Research Libraries Group (RLG), and the Public Record Office.

A series of Cedars Guides is designed to disseminate the achievements of the project in the following five major areas: Preservation Metadata; Intellectual Property Rights; Collection Management; Technical Strategies; and the Digital Archiving Prototype. The guides are available in printed form and are also available from the project website at <http://www.leeds.ac.uk/cedars/>.

5.1.2 DILIGENT

Within DILIGENT, the so-called “information objects” form the central entities of the storage manager. Information objects encapsulate storage units which are documents (represented as files or database-managed Binary Large Objects (BLOB) fields) including lists of associated plain metadata tags (basically key-value-pairs). Information objects may link to other information objects by means of role-attributed relationships. Those generalized inter-object relationships can be exploited to model various kinds of DILIGENT-relevant relations like (materialized) collections, archives, complex objects types, metadata associations, and notification relationships.

Each information object has a URI used to identify its storage location. The same identifier is used to identify the related metadata.

External content sources are accessible via “content wrappers”, which are part of the Content Management Service. External objects also have unique identifiers in DILIGENT. The wrappers map them to the external access policies and can transform or convert data formats and schemas. Available metadata on external data sources are migrated by the Metadata Broker to the DILIGENT metadata storage for easier access.

This information model is exposed to services through the Storage Management Service and a XML schema exists to map this internal model to a representation in the WS world. The Storage Management Service exposes each collection, document, and archive as a WS-Resource.

In the DILIGENT information model the Preservation Description Information (PDI) can be considered as an additional form of metadata conceptually stored along with the information object itself. No specific support is natively provided by the infrastructure concerning this kind of metadata even if ad-hoc services dealing with it can be plugged in.

5.1.3 SAFE

Current CCSDS Standards for Data Packaging have not undergone a major revision in 15 years. In that time, the computing environment and the understanding of metadata have changed radically:

- Physical media → Electronic Transfer





- The primary form of access to and delivery of, both archived and recently produced data products has shifted from hard media to include substantial network delivery.
- No standard language for metadata → XML
 - After 'bits' and 'American Standard Code for Information Interchange (ASCII)', the language 'XML' can be viewed as the next universal data standard, as it has grown exponentially.

Homogeneous Remote Procedure Call → Common Object Request Broker Architecture (CORBA), SOAP

- Communicating heterogeneous systems are increasingly using standard remote procedure calls or messaging protocols. The primary Remote Procedure Call (RPC) and messaging protocol for the WWW is SOAP, a XML based protocol.
- Little understanding of long-term preservation → OAIS RM
 - The **OAIS** Reference Model has become a widely adopted starting point for standardization addressing the preservation of digital information. The **OAIS** defines and situates within functional and conceptual frameworks the concepts of Information Packages for archiving (Archival Information Packages, or AIPs), producer submission to an archive (SIPs), and archives dissemination to consumers (Dissemination Information Packages, or DIPs).
- Record formats → Self describing data formats
 - Commensurate with XML, and rapidly growing computing power and storage capabilities has been an increasing tendency to use data formats that are more self-describing.

Further, there are a number of new requirements that are needed in the Space domain to facilitate such functions as being able to describe multiple encoding of a data object, and to better describe the relationships among a set of data objects. Therefore it is necessary to define a new set of packaging standards while maintaining the existing functionality.

Although XFDU is still in the pipeline of the CCSDS development, the maturity of the current working draft suffices for the primary needs identified for SAFE. SAFE has, therefore, been designed to fully comply with the current definition of this emerging standard, offering several advantages and opportunities in the following ways:

- XFDU inherits from the experiences gathered from the several international agencies, laboratories and companies composing the CCSDS;
- XFDU shares at least the same compliance level with the **OAIS** reference model;
- XFDU makes SAFE immediately compatible with the software due to be developed within or outside ESA scope;
- XFDU facilitates interchanges between several archiving management systems maintained by agencies following the CCSDS recommendations;
- SAFE supports the development of XFDU by providing the CCSDS working groups with returns from experience of using and implementing the intermediate working drafts, and;
- SAFE supports the development of XFDU by providing software material implementing the standard.





5.1.4 CIDOC Conceptual Reference Model (CRM)

The CIDOC Conceptual Reference Model (“CRM”) is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It is the culmination of more than a decade of standards development work by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Work began in 1996 under the auspices of the ICOM-CIDOC Documentation Standards Working Group. Since 2000, development of the CRM has been officially delegated by ICOM-CIDOC to the CIDOC CRM Special Interest Group, which collaborates with the ISO working group ISO/TC46/SC4/WG9 to bring the CRM to the form and status of an International Standard.

The primary role of the CRM is to enable information exchange and integration between heterogeneous sources of cultural heritage information. It aims at providing the semantic definitions and clarifications needed to transform disparate, localised information sources into a coherent global resource, be it within a larger institution, in intranets or on the Internet. Its perspective is supra-institutional and abstracted from any specific local context. This goal determines the constructs and level of detail of the CRM. More specifically, it defines and is restricted to the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal reference ontology.

It does not define the terminology that may typically appear as data in the respective data structures. Instead it foresees the characteristic relationships for its use. It does not aim at proposing what cultural institutions should document but explains the logic of what they actually document, and thus enables semantic interoperability.

The CRM intends to provide an optimal analysis of the intellectual structure of cultural documentation in logical terms. As such, it is not optimised to implementation-specific storage and processing aspects. Rather, it provides the means to understand the effects of such optimisations to the semantic accessibility of the respective contents.

Specifically, the CRM aims to support the following functionalities:

- To inform developers of information systems as a guide to good practice in conceptual modelling, in order to effectively structure and relate information assets of cultural documentation.
- To serve as a common language for domain experts and IT developers to formulate requirements and to agree on system functionalities with respect to the correct handling of cultural contents.
- To serve as a formal language for the identification of common information contents in different data formats; in particular to support the implementation of automatic data transformation algorithms from local to global data structures without loss of meaning. The latter being useful for data exchange, data migration from legacy systems, data information integration and mediation of heterogeneous sources.
- To support associative queries against integrated resources by providing a global model of the basic classes and their associations to formulate such queries.
- To resolve free text information into a formal logical form. It is believed that advanced natural language algorithms and case-specific heuristics may take advantage of this should it prove beneficial. The CRM is however not thought to be a means to replace scholarly text, rich in meaning, by logical forms, but only a means to identify related data.

Users of the CRM should be aware that the definition of data entry systems requires:- (1) support of community-specific terminology, (2) guidance to what should be documented and in which sequence, and (3) application-specific consistency controls.





By its very structure and formalism, the CRM is extensible and users are encouraged to create extensions to fulfill needs of more specialized communities and applications. The overall scope of the CIDOC CRM can be summarised in simple terms as the curated knowledge of museums.

However, a more detailed and useful definition can be articulated by defining both the *Intended Scope*, a broad and maximally-inclusive definition of general application principles, and the *Practical Scope*, which is expressed by the overall scope of a reference set of specific identifiable museum documentation standards and practices that the CRM aims to encompass, however restricted in its details to the limitations of the Intended Scope.

5.1.4.1 Intended scope

The Intended Scope of the CRM may be defined as all information required for the exchange and integration of heterogeneous scientific documentation of museum collections. This definition requires further elaboration:

- The term “scientific documentation” is intended to convey the requirement that the depth and quality of descriptive information that can be handled by the CRM should be sufficient for serious academic research. This does not mean that information intended for presentation to members of the general public is excluded, but rather that the CRM is intended to provide the level of detail and precision expected and required by museum professionals and researchers in the field.
- The term “museum collections” is intended to cover all types of material collected and displayed by museums and related institutions, as defined by ICOM . This includes collections, sites and monuments relating to fields such as social history, ethnography, archaeology, fine and applied arts, natural history, history of sciences and technology.
- The documentation of collections includes the detailed description of individual items within collections, groups of items and collections as a whole. The CRM is specifically intended to cover contextual information: the historical, geographical and theoretical background that gives museum collections much of their cultural significance and value.
- The exchange of relevant information with libraries and archives, and the harmonisation of the CRM with their models, falls within the Intended Scope of the CRM.
- Information required solely for the administration and management of cultural institutions, such as information relating to personnel, accounting, and visitor statistics, falls outside the Intended Scope of the CRM.

5.1.4.2 Practical scope

The Practical Scope of the CRM is expressed in terms of the current reference standards for museum documentation that have been used to guide and validate its development. The CRM covers the same domain of discourse as the union of these reference standards. This means that data correctly encoded according to any of these museum documentation standards can be expressed in a CRM-compatible form, without any loss of meaning.

Users intending to take advantage of the semantic interoperability offered by the CRM may want to make parts of their data structures compatible with the CRM. The respective parts should pertain either to (1) the associations by which users would like their data to be accessible in an integrated environment, or (2) the contents intended for transport to other environments. This will ensure that the meaning encoded by its structure is preserved in another target system.

In that sense, the CRM does not propose a complete match of user documentation structures with the CRM, nor that a user should implement all CRM concepts and associations. It is intended to (1) leave





room for extensions to capture the richness of cultural information or (2) allow for simplifications for reasons of economy.

The CRM is a means to interpret structured information in a way that enables large amounts of data to be transformed or mediated automatically. As a consequence, the CRM does not aim to resolve free text information into a formal logical form; this does not fall under the scope of compatibility considerations.

The CRM can foresee the associations to transport such information in relation to structured information. The CRM is a formal ontology, expressible in terms of logic or a suitable knowledge representation language. Its concepts can be instantiated as sets of statements that form models of the assumed reality referred to in a structured document. Any encoding of CRM instances in a formal language that preserves the relations to the CRM classes, properties and inheritance rules among them is regarded a “CRM-compatible form”.

A part of a documentation structure is compatible with the CRM, if a deterministic logical algorithm can be found, that transforms any data correctly encoded in this structure into a CRM-compatible form without loss of meaning. No assumptions are made about the nature of this algorithm. It may in particular draw on other formal ontologies expressing background knowledge such as thesauri. The algorithm itself can only be found and verified intellectually by understanding the meaning intended by the designer of the data structure and the CRM concepts. By the term “correctly encoded” we mean that the data are encoded so that the meaning intended by the designer of the data structure is correctly applied to the intended meaning of the data.

Information system implementers may choose to provide export facilities of selected data into a CRM-compatible form. They may further choose to provide a service to access selected data by querying with CRM concepts. It is not regarded a loss of compatibility, if certain subclasses and subproperties of the CRM are not supported in such a service. In that case it is regarded essential that the services publishes the set of CRM concepts it supports.

5.1.5 InterPARES

OAIS includes the main concepts relevant to the definition of authenticity, like content, fixity, provenance and context. However, these are not meaningful enough without a semantic approach able to provide the crucial/essential components for preserving resources. Without this effort and specifically if the required context is not fully represented, the digital resources risk being poorly represented. The semantics relevant to authenticity as developed by InterPARES (and more specifically detailed in the Annexe) are able to give the required contextual information and the necessary description for the digital objects with respect to their specific domain. These elements could be included in the PDI by using the appropriate tools i.e. ontologies, as well defined mechanisms able to share standardised information representations and dynamic relations. This is necessary to ensure a more detailed and specific implementation of OAIS components to qualify and make more efficient the use of OAIS model for specific designated communities. The relationships with the findings of InterPARES related to the authenticity could provide a useful integration in this crucial aspect of the preservation function.

5.2 REPRESENTATION INFORMATION

(A) PROJECTS AND OTHER MAJOR INITIATIVES

5.2.1 SAFE

A SAFE product, which can also be considered a “XFDU package”, wraps or references EO data and associates them with information expressed in EO vocabulary. The primary objective of SAFE is to





hold the Level-0 (L0) data, which is close to the telemetry level, but it has, moreover, been qualified for the packaging of higher levels products, i.e., the ENVISAT GOMOS. Level 1 and 2 products have already been successfully implemented in SAFE.

All SAFE products contain the following metadata:

- Acquisition period: the acquisition period metadata that provide the time extents of all the data contained in the SAFE product. It is mainly dedicated to allow a fast time ordering and framing of the overall contained data.
- Platform/Sensor identification: the platform identifies the system (satellite/aircraft) that acquired the EO data wrapped by the SAFE product. It has sub-elements that unequivocally identify the platform as well as the specific sensor that acquired the data.
- Product History: a processing log collecting the historical information dedicated to the maintenance and the trace ability of the product. The main feature of this logging system is its capability to store several processing threads regarding all the components that affected the product. As an example, if the product originates from the concatenation of several data objects, all logs of the involved objects will be kept, identifying their precise role of each in the production of the described SAFE product.

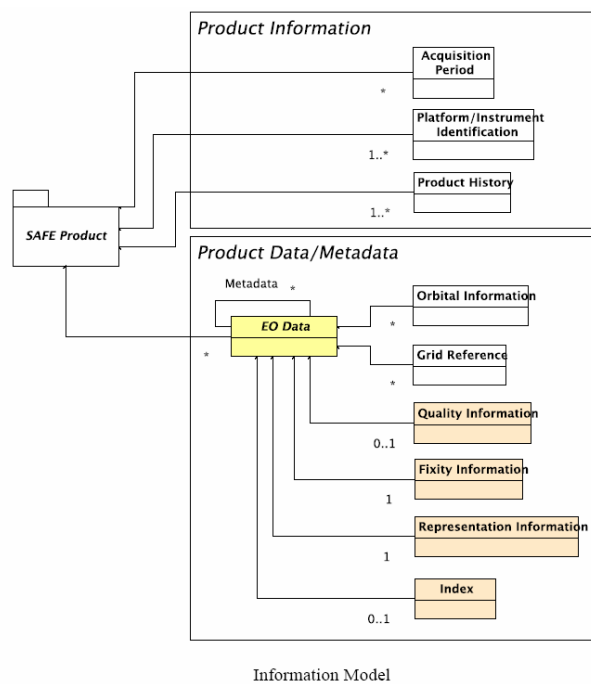


Figure 15 SAFE Information Model

For each wrapped or referenced EO dataset, a collection of metadata information may be attached:

- Orbital information: the reference to the trajectory of the platform that acquired the data. This information may locate one or several orbit paths, the corresponding cycle, track, etc.
- Geolocation information: the information locating the product footprint on the Earth's surface, either as a series of tie points or by reference to a world reference system of the acquiring platform. The Geolocation information, may also attach additional information to each localized element, including cloud coverage vote notation, meteorological information, etc.





- Quality/Fixity information: information about the quality of an EO dataset. SAFE makes use of techniques (i.e., eXtensible Markup Language Path Language (XPath)) that allows the precise location of the corrupted or missing elements up to the bit level.
- Representation information: any data contained in a SAFE product shall be accompanied with its representation information formally and numerically exploitable. Although the semantic information is partially implemented (i.e., all elements composing the dataset are named but the semantic links to standardized vocabulary is not embedded in the product), the SAFE aims at complying with the **OAIS** reference model in that area for assuring the maximal theoretical long-term preservation.

Finally, SAFE does not limit the information to the content listed above but supports extensions as far as they preserve the integrity of the mandatory items.

5.3 STRUCTURE INFORMATION

(A) PROJECTS AND OTHER MAJOR INITIATIVES

5.3.1 BRICKS

As mentioned in the previous sections, BRICKS provides different data management services, each own defining its own model. The hearth of the BRICKS information model is a DLObject, which is the basic entity for all the digital library operations. DLObjets are organized in hierarchical structures; each of them can contain several sub components which represent the structural and temporal dimensions of the content.

BRICKS DLObjets are stored as JCR Nodes, defined by the bricks:dlobjct type, which includes a set of mandatory attributes:

- owner
- persistent identifier
- parent collection
- parent node / institution
- version

Additional properties are stored for each BRICKS Collections, which are also derived from DLObject. Each DLObject Node contains an arbitrary set of additional properties and sub-nodes, defined and enforced through the JCR type system. Nodes can hold references expressed as URNs, or directly contain binary streams, which are stored into the available backends. Node properties can also be used to reference other Nodes, i.e. for guaranteeing referential integrity.

Additionally for every DLObject an arbitrary number of metadata Records are managed, depending on the ontologies loaded into the system.

5.3.2 DILIGENT

DILIGENT Physical and Logical Model

The physical model is implemented through a relational database. Every storage node hosts one database instance. Each database builds on an Entity-Relationship (E/R) schema, which represents the information object type (logical model description).

- Each information object may comprise a list of storage properties represented as simple key-type-value associations. Those storage properties are atomic whereas complex metadata (like indexes, multimedia features) may also be represented as separate information object related to the object.





- Each information object can be an abstraction of a file-based document, a BLOB-field representation of content, a simple document with no related content (besides storage properties and/or object references), and a complex object that references to other objects. The storage manager does not maintain consistency of the document content and explicit references to other objects.
- An object reference “links” two information objects. Each information object may (1) reference many other objects and (2) be referenced by many objects (m-n relationship). For instance, a reference type may be “indexes” with a role name that gives additional information, like “full-text index”. In addition, a positioning attribute helps in representing an object that references to other objects, like an aggregate made up of components that have to be fitted together in a certain order. Propagate role provides removal constraints to the storage manager. Different reference types may impose different rules when deleting a referenced object. For instance, component objects appearing as collection members will not be deleted if the collection is removed but any related index will be removed.

5.3.3 Fedora

The Fedora object model is defined in XML Schema language. The FOXML schema provides a complete expression of the Fedora object model. For more information, also see the Introduction to FOXML in the Fedora System Documentation. The Fedora object model also supports versioning for data streams and disseminators.

The basic components of a Fedora digital object are:

- PID: a persistent, unique identifier for the object.
- Object Properties: a set of system-defined descriptive properties that is necessary to manage and track the object in the repository.
- Data stream(s): The components in a Fedora object that represents MIME-typed content item. An object can have more than one Data stream. The content of a data stream can be either data or metadata, and this content can either be stored internally in the Fedora repository, or stored remotely (in which case Fedora holds a pointer to the content in the form of a URL). Every object has one DC metadata data stream by default.
- Disseminator(s): The components in a Fedora object that associates an external service with the object for the purpose of providing extensible views of the object or of its data stream content. An object can have zero or more Disseminators.

5.3.4 SAFE

SAFE Physical Model

A SAFE product physically contains the following components:

- Manifest file: a XML document comprising the XFDU Manifest file. It contains the definition of the Information Package Map, the wrapped Metadata Objects (i.e., in general all Metadata Objects are embedded in the SAFE Manifest file), the wrapped Data Objects (i.e., Data Objects are rarely embedded in the SAFE Manifest file) and references to the external files containing the Metadata and Data Objects.
- Binary or XML files: the data or metadata object contents. Currently, only two types of files have been identified, i.e., binary matching MIME octet stream definition and XML





documents. Each of these files shall be accompanied with one or more XML Schema document controlling its content.

- XML Schema files: the representation information of the data held by a SAFE product. In comparison to XFDU, SAFE does not allow multiple notations for storing the representation information of its objects. This restriction is mainly imposed because SAFE does not only reference a representation information technique but intends to define it. In order to represent the binary information, SAFE also defines specific markups that annotate the XML Schema documents to provide information on the physical structure, i.e., the so-called Structured Data Format (SDF) markups. Thanks to these specific annotations, the contents of the binary files are described up to the bit level with a common technique as for XML documents.

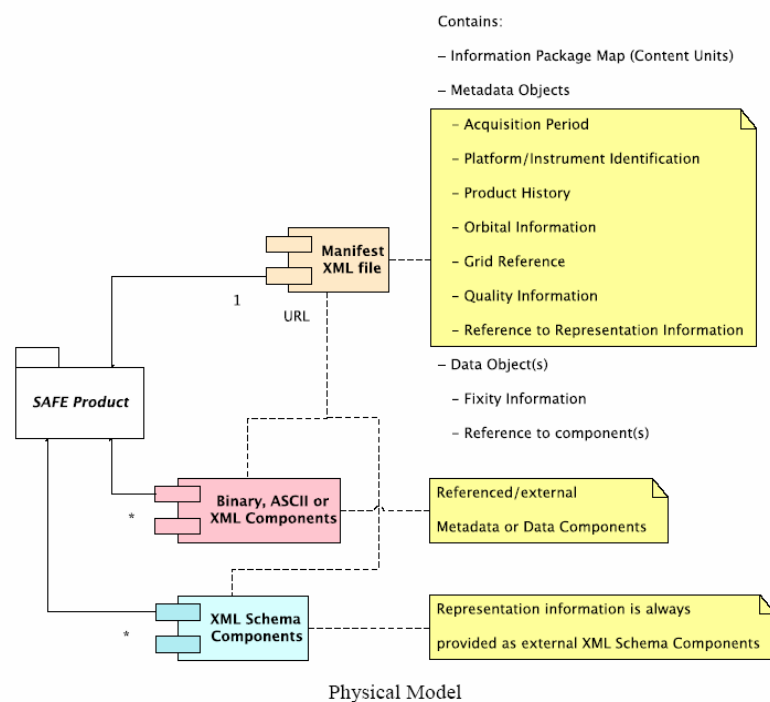


Figure 16 SAFE Physical Model

(B) OTHER TECHNOLOGIES

5.3.5 Data Structure Description Languages

The purpose of a Data Structure Description Language (DSDL) is to map the data bits within a data stream/file to data values such as numbers and strings and then show how these values are ordered with respect to one another within the data hierarchy. They provide one type of **OAIS** structure representation information. To define the structure of a data stream/file the languages must provide the following:

1. A way of describing the structure of the individual bits and how they map to atomic types, such as integers, reals, characters, arrays and strings. In effect this virtualises the basic data access routines that are normally part of the computer system and programming libraries. This is sometimes called the physical description.
2. A way of describe the hierarchical structure of the values in the data stream/file. This is sometimes called the logical description.





The current data structure description languages can be roughly grouped into two types:

- 1 Languages that describe the hierarchical structure of a given data file and the atomic types that make up that hierarchy. A good example is XML schema for hierarchical XML documents.
- 2 Languages that are really just a subset of a full computer programming language. The subset of features usually includes a limited definition of atomic types plus the ability to loop and branch. They also usually include a set of IO functions.

Currently no one single language provides all the capabilities necessary to describe arbitrary hierarchical data in text, XML or binary format. The first type are the most promising for the use in **CASPAR** as part of the implementation of the simple/complex object virtualisation process. But describing the structure of data is only one half of the picture. To support the full virtualisation process and add meaning to the values extracted from the data, then the semantics that is associated with the data values and data hierarchy must also be included, as already indicated in **OAIS**.

There are several DSDLs:

1. Enhanced Ada Subse T (EAST) - ISO 15889-2000,
2. XML Schema and related technologies – W3C Standard,
3. Data Request Broker (DRB) – part of the SAFE format.
4. BinX - <http://www.edikt.org/binx/>
5. [Data Format Description Language \(DFDL\)](#)
6. Flavor (Formal Language for Audio-Visual Object Representation) - <http://flavor.sourceforge.net/>

EAST provides a complete set of features for providing a physical description but lacks some features necessary to logically describe some of the more complex data formats. EAST can deal with both binary and text formats. There is an API (closed source) for EAST but maintenance and support for it are poor. There are two tools (closed source) for producing EAST descriptions which are OASIS (<http://east.cnes.fr>) and the DEBAT BEST Modelling tools (<http://debat.c-s.fr>), but again the support and ongoing maintenance for these tools is questionable.

XML Schema is currently limited to describing XML data in text format. There are a wide verity of tools for producing XML Schema and accessing XML data.

DRB extends XML Schema (logical description) to binary formats by adding non-standard extension to XML Schema and XQuery/XPath. The inclusion of XQuery allows DRB to provide a rich set of features for producing a logical description but at the cost of increasing the language complexity unnecessarily. DRB provides few features for producing a physical description, and is essentially limited to byte-order and IEEE 754 floating-point formats. There is a current and maintained API (<http://www.gael.fr/drj/site/> closed source) for DRB with user support. To produce descriptions you can use any of the tools that are used for defining an XML Schema.

BinX provides the ability to describe binary data in continuous stream, it cannot describe data that uses offset pointers to locate data structures within data. Thus its logical description capabilities are limited. It lacks the ability to give a full physical description, and just provides a fixed set of data types. It also does not allow the description of structured text or XML data. These facts currently limit the use of BinX to very simple data structures. BinX defines its own data description language, storing the structure information in an XML document. The language does have an XML schema, which does make it easy to use the information contained in the descriptions without a specific API. BinX also provides tools for generating the descriptions as well as and API for using them. The last





release was in 2005, which suggests that development may have stopped, but in the development plans it is suggested that in the next release the ability to deal with structured text data will be added.

DFDL is very similar to DRB. It uses XML Schema to describe the logical structure of the data and then adds extensions to XML Schema to add the physical description of data types. It uses a reduced XPath scheme to point to access data values with the description. As with DRB its physical description capabilities are limited to byte-order and IEEE 754 floating-point formats. DFDL is still in the specification stage, so no APIs are provided yet to use the descriptions with. It is possible to produce descriptions with existing XML schema definition tools such as XMLSpy and the current specification (v1.0-19 <http://forge.gridforum.org/sf/go/doc14380?nav=1>) is very detailed.

Flavor uses the reduced programming language approach to data structure descriptions. It provides a simple programming language with data types, loops and branches. Using the language you can describe most continuous data streams, but as a lot of the data description languages, it cannot deal with offsets to structures with data. The data types are also limited. The purpose of Flavor was to provide formal multimedia data stream descriptions within the MPEG 4 standard. Flavor provides tools for converting the description into code such as Java or C++. This code can and then be used in an application to access the data streams. This is a fundamentally different way of using the data descriptions, and is less general than providing an API specification as most other languages do. A tool to convert data to XML using the description is also provided. No tools are provided to produce the descriptions, other than using a simple text editor to create them.

5.4 SEMANTIC INFORMATION

(A) PROJECTS AND OTHER MAJOR INITIATIVES

5.4.1 BRICKS

The BRICKS Information model, where DObjects play the role of the central information entity, is suitable for the implementation of any semantic information structure, as DObject metadata records are already defined, managed and searched through user-defined ontologies. Thanks to the object referencing mechanism, the hierarchical organization of DObjects can be easily extended, through search views and other approaches in order to model semantic relationships among DObjects.

5.4.2 Fedora

Fedora defines a generic Digital Object Model that can be used to express many kinds of objects including documents, images, electronic books, multi-media learning objects, datasets, metadata, and many other entities. Fedora supports aggregation of one or more content items in a digital object.

5.4.3 SAFE

As specified in the XFDU model, a SAFE product is a logical tree of "Content Units" [2] forming the so-called "Information Package Map". Conversely to XFDU, only one map is expected per SAFE product. The root Content Unit has predefined associations to the information applicable to the overall product, i.e., at least the "Acquisition Period", the "Platform/Sensor Identification" and the "Product History". The structure of the children Content Units is less constrained and depends mainly of the logical view of the wrapped data. In most cases, one Content Unit matches one EO dataset and its accompanying metadata. Several Content Units may, however, share the same metadata.



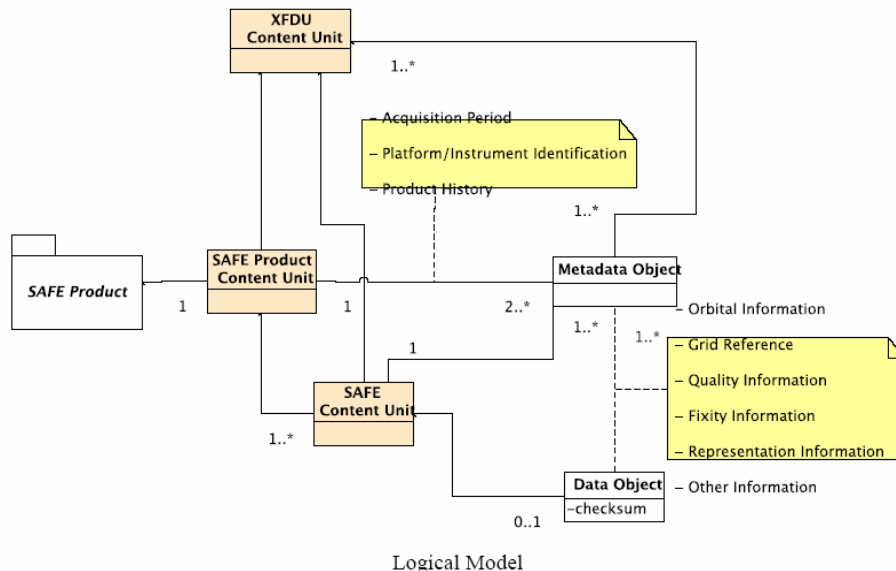


Figure 17 SAFE Logical Model

5.5 OTHER INFORMATION

(A) PROJECTS AND OTHER MAJOR INITIATIVES

5.5.1 BRICKS

BRICKS Framework supports a decentralized architecture, where each BRICKS Node (BNode) contributes its services to a peer-to-peer network. A BNode could be seen as a set of services that are required to manage an institution's presence in the system, and to provide services for the rest of the community. A BNode consists of three types of components: fundamental, core, and basic Bricks. Most of them are standard Web services, described by WSDL documents and registered with an UDDI compatible repository used also for discovering appropriate services.

New bricks may be added without affecting the rest of the system, providing new functionality or reusing the existing component for exposing new interfaces.

5.5.2 DILIGENT

DILIGENT is based on a Services Oriented Architecture (SOA). The services are organized into three layers, namely the Collective Layer, the Digital Library Layer and the Application Layer.

- **Collective Layer:** it contains basic services such as Security Management, Service Configuration and Registry.
- **Digital Library Layer:** it contains specific services for Digital Libraries (DLs) such as Process Management, Content & Metadata Management, Index and Search Management.
- **Application Layer:** it represents the presentation layer and is based on the Java Portlet Specification V1.0 developed under the Java Community Process as JSR 168.

The DILIGENT infrastructure supports the addition of new services and the combination of services in user defined workflows, thus potentially any technique implemented via WS as well as unit of processing (jobs) can be plugged into the system.





5.5.3 SAFE

The SAFE specifications provide abstract definitions and ruling for handling a product. Similarly to XFDU, SAFE offers a framework that assures the consistency between all products but may not suffice for the complete definition of a product. Actually, EO products are generally accompanied with metadata specific to the mission or the sensor that acquired the referenced data. Moreover, a specific product may control more precisely the content of the generic information such as the Information Package Map structure, the platform name, etc.

The specifications of SAFE have therefore been broken down in two main layers:

- the core level that controls the relationship with XFDU and the general structure that shall be followed by all SAFE products;
- the specialization level that implements the core level up to obtain a complete and accurate definition of a specific product type.

5.5.4 Computer Emulation and Virtualisation Technologies

It is true that software plays an important role in preserving data and its information content. Software typically provides a means for accessing, processing and rendering data and information in accordance with algorithms that are specific to a particular domain. As a result, software captures a lot of domain specific knowledge related to a particular data set that may be difficult to recover via other means (publication, books etc) and re-implement in software in the future. For this reason, the ability to run existing software in the future would aid the preservation process greatly. Solutions to the software preservation problem are likely to include computer emulation and virtualisation technologies.

We shall use the definition of computer emulation to coincide with the definition of an "instruction simulator"²³. The result is an emulator that can run the system software (operating system and some device drivers) and all the applications which run under that operating system. The host that the emulator runs on does not have to have the same instruction set or hardware as that being emulated.

A wide variety of emulators currently exist which successfully run current and past operating systems and application software, some good examples are,

- 3 QEMU - Which emulates a variety of systems including ARM, SPARC, PowerPC, MIPS, x86 and x86-64 instruction sets. It allows the running of full operating systems (Windows, Linux and Mac OSX etc) and applications²⁴.
- 4 Bochs IA-32 Emulator Project - emulates x86 and x86-64 instruction sets but is highly portable and runs on a variety of host systems such as SPARC and PowerPC²⁵.

So there is no question that emulation of current and past systems can be done and done in a portable way. The problem with the current set of emulators is that the information on the instruction sets and hardware is locked away in the emulator source code. Although there is typically documentation on system instruction sets and hardware, they have a habit of being incomplete, and a certain amount of knowledge about the systems can be lost over time, making it practically impossible for someone in the future to create an emulator.

²³ "Preserving Computing's Past: Restoration and Simulation", by Max Burnet and Bob Supnik, from the Digital Technical Journal, Volume 8, Number 3, 1996.

²⁴ QEMU – <http://fabrice.bellard.free.fr/qemu/>

²⁵ Bochs – <http://bochs.sourceforge.net/>





The other problem with existing emulators is that they do not guarantee that the "Look and Feel" of application software is reproduced accurately, they only guarantee that the software will run. "Look and Feel" is important for certain types of software and data, including documents and games, but it can present problems when running any type of software. What is required for "feel" is an intermediate step between full system simulation⁹ and emulation, i.e. taking into account the "average timing" of instruction execution on a real system within the emulator. Some work on instruction timing has been done as a follow-on from research into embedded system power requirements²⁶. Measuring, recording and using instruction timing information within emulators to reproduce "feel" is a potential research topic. The "look" really applies to emulation of the systems display and colour rendering on that display. Currently it is not a problem as most system displays operate in a similar manner, but in the future a full description of how to render the information from hardware graphics device would be required.

5.5.4.1 Virtual Machines

Virtual machines are not emulators, and the term virtual machine is used in a variety of contexts.

The VMware virtual machine²⁷ emulates all the system hardware (graphics, disk etc) but does not emulate the CPU instructions. As a result, VMware will only run x86/x86-64 operating systems and applications on a x86/x86-64 host system.

QEMU also includes a x86 virtualisation module for speed improvements, as you do not need to emulate the CPU if the emulator is running on x86 system (host) and running x86 operating systems. The advantage of QEMU is that it is open source.

The Java virtual machine²⁸ and the .NET CLI²⁹ use the principle of defining a completely new instruction sets that can be implemented in software (virtual machine) on any host system. Typically they run a particular language (Java or C#) which is compiled first to byte code and then to the host system instruction set via the virtual machine. It is possible to implement other languages for the virtual machines, but typically this still involves porting the application source code to the virtual machines due to the fact that they also defined class libraries as part of their specifications. As a result they do not provide a general solution to software preservation without a significant amount of programming effort. Another problem with the virtual machines is similar to that of emulation, instruction timing is undefined.

5.5.4.2 Universal Virtual Computer (UVC)

Raymond Lorie proposed the UVC³⁰ concept in 2000 in a research paper written for IBM. It was later published more widely in an article in RLG DigiNews.³¹ The UVC approach relies partially on emulation concepts and aims to allow digital objects to be retained in their original format alongside a program, which can decode the data and present it in an understandable form.

In brief, the UVC is a simple virtual computer architecture that will run on any existing hardware platform. The UVC is a computer in its functionality; it is virtual because it will never have to be built

²⁶ *Modeling assembly instruction timing in superscalar architectures*, Beltrame, G. Brandolese, C. Fornaciari, W. Salice, F. Sciuto, D. Trianni, V. CEFRIEL Res. Centre, Milan, Italy. System Synthesis, 2002. 15th International Symposium on.

²⁷ VMware virtual infrastructure software – <http://www.vmware.com/>

²⁸ The Java Virtual Machine Specification, Second Edition <http://java.sun.com/docs/books/jvms/>.

²⁹ Common Language Infrastructure Standards (CLI) - <http://msdn2.microsoft.com/en-us/netframework/aa569283.aspx> - ISO/IEC 23270 (C#), ISO/IEC 23271 (CLI) and ISO/IEC 23272 (CLI TR).

³⁰ See http://en.wikipedia.org/wiki/Universal_Virtual_Computer

³¹ R.A.. Lorie, "A Project on Preservation of Digital Data", *RLG DigiNews*, vol. 5 no. 3. At <http://www.rlg.org/legacy/preserv/diginews/diginews5-3.html> and *The UVC: a method for preserving digital documents*, http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf.





physically; and it is universal because its definition is very basic and is guaranteed to remain unchanged in present and future.

The UVC architecture relies on concepts that have existed since the beginning of the computer era: - memory, registers, and a set of low-level instructions. The fact that the computer is virtual and that performance is of secondary importance allows for a simpler, more logical, perhaps less optimised, design.

A UVC interpreter, a program compiled to (or written in) the machine language of UVC, is completely independent of the architecture of the computer on which it runs. To access data in the future on a given platform, a UVC interpreter should be written for the hardware/software configuration on which the data will be accessed.

Note that the UVC application will not allow reinterpretation of the preserved data, once the data and the decoding program are preserved. A new representation for the data can't be generated in the future.

This UVC is utilized in the Digital Information Archiving System (DIAS) solution, developed by Koninklijke Bibliotheek (KB) and IBM, where a test implementation has been developed for preserving digital images.³²

The UVC consists of a virtual machine in a similar way to Java and the .NET CLI does. But in addition to this it includes data structure and semantic descriptions combined with decoding software defined for the UVC. The result is a program which decodes and displays a given set of data in a particular format. The disadvantage of this approach is that you have to re-implement software for the UVC architecture, which is time consuming and expensive, particularly for complex scientific software. It does however have applications in the areas of document and image preservation where only a handful of decoding programs would have to be written to cover a large number of data sets and archives. Some work has been proposed on combining the UVC and current system emulation technologies, which may increase the usefulness of the UVC.

³² See http://www.kb.nl/hrd/dd/dd_onderzoek/uvc_voor_images-en.html.





6 OAIS PRESERVATION DESCRIPTION INFORMATION (PDI)

6.1 GENERAL PRESERVATION DESCRIPTION INFORMATION

(A) PROJECTS AND OTHER MAJOR INITIATIVES

6.1.1 Cedars

The hierarchical structure of the Cedars specification demonstrates its basic dependence upon the OAIS information model. The first three levels of the hierarchy inherit the exact terminology and some of the definitions used in the OAIS model.

The *Reference Information* section of the PDI has elements for a 'Resource Description' and a placeholder for any 'Existing Metadata'. The Cedars specification does not make any specific recommendations as to which elements would be included in the 'Resource Description,' but notes that any project-specific implementation would use an instantiation of the Dublin Core Metadata Element Set (DCMES). In a similar way, the precise way in which 'Existing Metadata' would be stored or utilized is not defined. In an operational repository, it is possible that at least some of the Descriptive Information and Reference Information would be generated automatically or extracted from metadata that already exists, e.g. in publishers databases or library catalogues.

Context Information has one sub-element, referring to 'Related Information Objects.' This is supposed to specify information objects that are judged to have a significant relationship to the object being preserved. Again, what precise information would be required (e.g. an identifier, descriptive information, etc.) is not defined.

Provenance Information makes up the largest part of the Cedars metadata specification. The 'History of Origin' sub-section is intended to record the reasons why the object being preserved was created, its custody history before ingest, and to document why it is being preserved. This section also records technical information about the original technical environment of the object and any prerequisites with regard to software, operating systems, etc. A separate section on 'Management History' is supposed to keep information about the ingest process, and the policies and actions applied to objects since they were added to the repository. A final section on 'Rights Management' comprises a detailed set of sub-elements to help record and manage the intellectual property rights held in objects.

Fixity Information contains a single sub-element, 'Authentication Indicator,' which is intended to record mechanisms used to ensure the digital object's authenticity, e.g. digital certificates or a checksum.

6.1.2 BRICKS

The ability to manage descriptive metadata in RDF using arbitrary schemas expressed through OWL ontologies, makes the representation of PDI an easy task for BRICKS Metadata management services. The creation of a specific OWL ontology is the only step needed, The Digital License management services may also be used for storing some information concerning right-holders, provenance, fixity.

6.1.3 Fedora

Fedora manages PDI as part of its Core services, undertaking the creation, modification, and deletion of digital object metadata. Much of Fedora's ability to support the proper administration of PDI will depend on a Preservation System manager's configuration of metadata. This metadata will need to properly capture and manage the appropriate provenance, reference, fixity, and contextual information.





6.1.4 MUSTICA

6.1.4.1 Dublin Core (DC)

As basic as it might seem, we recall here that DC is a widely used standard, which recurs ubiquitously in almost all of the standards we encountered during this study. DC elements are identified by persistent URLs (PURLs), their semantics are well defined and bindings exist towards most common metadata syntaxes such as XML Schema and RDF. The DC elements are in general applicable to any kind of content.

6.1.4.2 The MUSTICA data model for musical works

The MUSTICA data model is formalized as a set of XML Schemas, which means that any MUSTICA description can be encoded in XML. The main entities in the model are works. A work has a set of metadata describing title, provenance and other descriptive and preservation information. Multiple versions of a work can exist. A version describes the score, other instructions, equipment and instruments needed for a performance. All digital assets are provided by the producer (composer or musical assistant) in one or more archives (tar-gzip, bzip2 or ISO 9660 images) and should be referenced by the description. For example an instruction description can point to a technical scheme, a score description to a directory containing the score files etc.

Executions of each version can also be registered in MUSTICA, with descriptors about contributors, equipment, time and place, and various media files for audio and video recordings.

For detailed information about the MUSTICA project, please refer to Annex VI – The MUSTICA Project.

6.1.5 CIDOC Conceptual Reference Model

CIDOC CRM is an ontology and is not strictly applicable to the OAIS Information Model. That said, it can be used to represent archival concepts, or rather to conceptualize the specific profiles related to provenance, fixity, reference and context. Many of its properties relate to creation, identification, alteration, temporal and spatial qualification, etc. It is crucial to determine to what extent adopting the model: CIDOC CRM can be extended adding entities and properties thanks to Open World Assumption.

It may seem a good strategy to represent your own domain, or at least the most specific profiles, using your own concepts and then merge them into CIDOC CRM in order to have a perfectly “customized” CIDOC CRM-based model. Alternatively, this strategy may not recognize that generic and pre-defined concepts are adequate to represent our world. It is important to find an equilibrium between the need for customization and the opportunity to adopt a given “vocabulary”.

6.1.5.1 Provenance Information

CIDOC-CRM can capture provenance information.³³ It includes (but it is not limited to):

- Name of creator(s);
- Administrative/Biographical history;
- Archival history; and
- Immediate source of acquisition or transfer.

³³ More detailed information can be found within the following document: *CIDOC CRM and OAIS Provenance*, <http://dev.dcc.ac.uk/CASPAR/bin/view/Main/CIDOC CRMAndOAIS-provenance>





6.1.5.2 Fixity Information

Fixity is related to Provenance (discussed above).

6.1.5.3 Reference Information (including for example Persistent ID)

CIDOC CRM include reference information. These elements are not strictly in keeping with the definition provided in the questionnaire but in the **OAIS** Model the bibliographic description is given as “reference information” for the digital library collection case. As a result the following reference information is included (but it is not limited to):

- Title;
- Dates;
- Level of description; and
- Extent and medium.

6.1.5.4 Context

CIDOC CRM could be used as the conceptual backbone for connecting all objects. In that case Context would be supported and includes (but it is not limited to):

- Scope and content.

6.1.5.5 Other PDI

CIDOC CRM does include the following additional useful data:

- Appraisal, destruction and scheduling information;
- Accruals;
- System of arrangement;
- Conditions governing access;
- Conditions governing reproduction;
- Language/scripts of material;
- Physical characteristics and technical requirements;
- Finding aids;
- Existence and location of originals;
- Existence and location of copies;
- Related units of description;
- Publication note;
- Note;
- Rules or Conventions; and
- Dates of description.

6.2 PDI FOR VIRTUAL ARTWORKS

6.2.1 Archiving the Avant-Garde

In preservation of art works, the Archiving the Avant-Garde and its related project, the “Preserving the Rhizome ArtBase”, also look at standards of metadata for cataloguing and documenting art works for preservation. The Avant-Garde proposes to develop new metadata standards for variable media art





works, which will be compatible and interoperable with other standards in arts and cultural heritage domains to enable consistent integration of information and long-term preservation of the new standards. In the “Preserving the Rhizome ArtBase” projects, two types of metadata were proposed to capture information about art works: (1) metadata about the original artwork and (2) metadata about the original software and technologies needed to run the work. Each of these includes:

- Descriptive metadata to be used for discovery and display of artwork;
- Administrative metadata to support the management of artwork in the archive, including rights to access and reproduction of the works, locations of related data files in archive, etc, and;
- Technical metadata used to describe technical aspects of the work, such as technologies used to run the work, programming languages, etc.

6.2.2 The Database of Virtual Art

The Database of Virtual Art³⁴ project in Germany is aimed at archiving expanded documentation about virtual artworks. The expanded documentation includes biographical and bibliographical information about the artists, images of installation structures, audio and video documentations, etc. The goal is that the archived documentation should be enough for performing analyses on the archived artworks. The basic approach taken by this database for preservation of artworks is to store as much information related to the works as possible, for future interpretation.

6.3 PDI-REFERENCE

There are large numbers of identifier systems for which persistence is claimed. New ones appear every few months. They have a variety of aims in terms of scalability, actionability and central control. The DCC Workshop on Persistent Identifiers³⁵ and the Erpanet workshop on Persistent Identifiers³⁶ provide a useful review of some of these systems. The PADI³⁷ web site also collects information on such systems.

All systems require some kind of resolver service to give a more specific location for the resource given the identifier. The persistence of this resolver is the key concern from the point of view of a preservation system.

Some systems recognise the need for a contractual obligation for cross-support, some systems are scalable, some are aimed specifically at collections held at digital libraries and most systems are not applicable to dynamic data such as that contained in data bases.

6.3.1 Persistent identification with the handle system

The handle system has been developed by the CNRI. It provides, by the means of a very scalable and well-designed distributed architecture, an efficient way to manage and resolve global persistent identifiers. An organization can set up a handle service obtaining a prefix from the handle system authority and running one or more handle servers. The prefix will be registered in the Global Handle Registry (GHR) mapping to the current address of the corresponding handle service. A handle service can be replicated in many handle sites, and any handle site can cluster an unlimited number of handle servers.

³⁴ h. O. Grau, "The database of virtual art for an expanded concept of documentation," <http://www2.hu-berlin.de/grau/database.htm>

³⁵ <http://www.dcc.ac.uk/events/pi-2005/>

³⁶ <http://www.erpanet.org/events/2004/cork/>

³⁷ <http://www.nla.gov.au/padi/topics/36.html>





A handle is not just a map from an identifier to an URL (as PURLs are, for example) but can have several typed data fields, some of which are used for administration. One task they fulfil is the permission management on the handle itself that is driven by a public key / private key certification mechanism. The single data-fields can have an access policy each as well.

6.3.2 Digital Object Identifiers (DOI)

An example of a handle system is the Digital Object Identifier (DOI) initiative. The International DOI Foundation, a non-profit organization, maintains the DOI system. This Foundation coordinates Registration Authorities (RAs) and developments around the DOI standard. DOIs are handles that allow RAs around the world to assign unique object identifiers. The DOI uses the Handle System metadata feature, and has developed a core metadata model that is shared among all RAs, including title, URL, etc.

6.3.3 URI etc

W3C defines a system of Uniform Resource Identifiers (URIs) of standard, persistent and unique identifier for digital resources on the Internet. The URI system includes Uniform Resource Name (URN) and URLs and for all of these a resolver service is required.

Uniform Resource Characteristics (URCs) are metadata encoded information about the resources.

All URNs include:

- Namespace Identifier (NID) code - indicates the identification system being used for the URN and facilitates the interpretation of the NSS.
- Namespace Specific String (NSS) – provides the local code that identifies the individual document (see IETF:RFC 1737³⁸, Functional requirements for Uniform Resource Names; IETF: RFC 2141³⁹, URN Syntax).

The international ISBN and ISSN agencies are registering URNs using 'ISBN' and 'ISSN' as the NIDs. A URN based on National Bibliography Numbers (NBNs) with 'NBN' as the NID has been registered and adopted by the Nordic Metadata Projects.

6.3.4 Persistent URL (PURL)

The Persistent Uniform Resource Locator⁴⁰ (PURL) was developed and implemented by the Online Computer Library Center Inc. (OCLC) as a naming and resolution service for general Internet resources. It is intended as an interim system to be used until the URN framework is well established. A PURL looks just like a URL, except it points to a resolution service instead of the actual location of the digital resource. The resolution service then redirects the user to the appropriate URL pointing to the current location of the particular resource.

Further information about PURLs is available at OCLC's Persistent URL Home Page. This site includes an overview of the PURL service, FAQs and details of the PURL-L mailing list. Organisations wishing to set up their own PURL resolver service can download the PURL software from this site.

6.3.5 ARK (Archival Resource Key)

The ARK⁴¹ (Archival Resource Key) scheme was developed by John Kunze as part of work undertaken for the US National Library of Medicine. The scheme required:

³⁸ <http://www.ietf.org/rfc/rfc1737.txt>

³⁹ <http://www.ietf.org/rfc/rfc2141.txt>

⁴⁰ <http://purl.oclc.org/>

⁴¹ <http://www.cdlib.org/inside/diglib/ark/>





- a link from the object to a promise for stewardship;
- a link from the object to metadata which describes it;
- a link to the object itself (or appropriate substitute).

6.3.6 Life Sciences Identifiers

The Life Science Identifiers⁴² scheme is now an OMG standard which was designed as standardized naming schema for biological entities in the Life Sciences domains, however it is much more broadly applicable.

6.3.7 Name to Thing (N2T)

Name to Thing⁴³ is another system proposed by John Kunze, and consists of a persistent identifier resolver and a consortium of cultural memory organizations.

⁴² <http://www.omg.org/docs/dtc/04-05-01.pdf>

⁴³ <http://n2t.info/>





6.4 OAIS PACKAGING

(A) PROJECTS AND OTHER MAJOR INITIATIVES

6.4.1 The Metadata Encoding and Transmission Standard (METS)

METS⁴⁴ is an open extensible XML schema standard developed by the library community providing the means to convey Representation Information necessary for the management of digital objects within an archive or repository and the exchange of objects between repositories or repositories and consumers. METS is based on the ‘common object format’ designed to allow the sharing of digital information and services for facilitating the interoperable exchange of digital materials between archives. METS schema was created in 2001 by the Digital Library Federation and developed and supported by the US Library of Congress. METS has been designed to solve the problem of keeping track of the relationships between and the storing complex and possibly geographically distributed representation information about digital objects for the long term and provide mechanisms and services for the efficient transfer and migration of the packaged digital object. METS is intended to be a tightly defined and structured container for a data object and its associated Representation information but allowing the flexibility to package together heterogeneous types of Representation information.

The METS schema is split into seven sections:

- **METS header** – Descriptive Information about METS object.
- **Descriptive Metadata** – References multiple instances of internal or external descriptive repInfo, such as File Structure Description and semantics.
- **Administrative Metadata** – would reference Provenance, PDI information and DRM repInfo elements.
- **File Section** – references all files comprising the digital object.
- **Structural Map** – Hierarchical structure of the digital object, containing links to content information, representation information and internal or external physical file locations.
- **Structural Links** – refers to hyperlinks between items referenced in the structure map.
- **Behavioural** – References repInfo about how content information is rendered to the user, such as software packages.

Using METS implementation for **CASPAR** and Long Term Digital Preservation could potentially provide a flexible mechanism for managing and referencing representation information, METS’s XML implementation⁴⁵ means it is platform independent and a good medium for file and data interchange allowing validation by DTD, XML schema or Schematron rules and incorporation with other XML schemas like Dublin core. Currently it is noted that the lack of software tools (though there is a Java toolkit⁴⁶) and support and the size of the standard have put many institutions off adoption of using METS. There is also a lack of built in support for **OAIS** concepts⁴⁷.

METS could be used to implement Information packages as described by the **OAIS** reference model and depending on its use, could have a role as Submission Information package, Archival information package or a Dissemination information package but the lack of standard software tool may hinder its adoption and use by **CASPAR**.

⁴⁴ <http://www.loc.gov/standards/mets/>

⁴⁵ <http://www.loc.gov/standards/mets/mets.xsd> – METS XML schema

⁴⁶ <http://hul.harvard.edu/mets/> – METS JAVA toolkit

⁴⁷ http://www.jisc.ac.uk/uploaded_documents/OAISmets.pdf – discussion on METS use with **OAIS**





6.4.2 XML Formatted Data Unit (XFDU)

As mentioned within SAFE project details, XFDU, a recommended standard developed by CCSDS⁴⁸, has been designed for the specific purpose to facilitate information transfer and archiving and is a technique for packaging data objects and representation information into a single packaged unit. XFDU is a physical container such as a ZIP file containing an XML manifest document tightly defined by an XML schema specified in the manifest, the manifest contains all the valuable information about the data inside the container and references to information outside the container⁴⁹. The manifest is split into four parts:

The Information package Map containing a logical view of the package, it a hierarchical xml tree representation of the package content. Each Node is a content unit and can be referred to from the other parts of the package by its content unit reference ID. The Data Object Section contains all physical information need to access the data objects. The metadata section records all the representation information for all items within the package. The Behaviour section associates executable code with the content of the package.

The Content Unit provides the primary view into the package as it refers to each of the data objects and it associates appropriate Representation Information (RepInfo) with each data object. The content unit reference to the RepInfo is via one or more metadata category pointers. For each such pointer there is a set of metadata classes that may be included in the manifest file or referenced by a URI. A content Unit may also reference external XFDU packages.

XFDU schema allows the package designer to define any metadata model by providing attributes for both the metadata categories discussed and a classification scheme for finer definition with categories. The schema “provides predefined metadata categories and classes via enumerated attributes that follow the **OAIS** information model”

Descriptive Information intended for the Finding Aids such as catalogs or Search Engines may be categorised as ‘DMD’ and further classified as ‘DESCRIPTION’ or ‘OTHER’.

Representation Information may be categorised as ‘REP’ and further classified as ‘SYNTAX’, ‘DED’ or ‘OTHER’

Preservation Description Information may be categorized as ‘PDI’ and further classified as ‘REFERENCE’, ‘CONTEXT’, ‘PROVENANCE’, ‘FIXITY’ or ‘OTHER’.

OASIS defines 3 different Information packages, SIP – Submission Information Package, a package sent from data producer for submission into the Archive, AIP – Archive Information Package, the actual packages stored by the archive and the DIP Dissemination Information Package, a package distributed to the consumer from the archive.

XFDU in the role of a SIP will provide support for linking Representation Information to data objects, An XFDU package being sent from producer to Archive for ingestion will contain all links necessary between the data and the representation information. If a syntactic XML schema description of the data is provided in the XFDU an automated validation of the packaged data could be performed, an XFDU manifest file can reference multiple data files which could be concatenated into one file on ingestion, thus providing support for handling large file collections. XFDU has an order attribute, which allows conditional processing making it possible for the archive to get a hierarchical and logical view of the incoming packages and launch a specific action on receipt of a specific package in the order. XFDU is described by a W3C schema, which can be extended to provide support for specialised content making it simple to add attribute and still have a valid manifest file checkable from a schema.

XFDU used in the role of AIP provides easy processing or transformation of the XML manifest file XML access languages (XPath, XQuery, XSLT) for many possible applications.

⁴⁸ <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206610R1/Attachments/661x0r1.pdf>

⁴⁹ <http://sindbad.gsfc.nasa.gov/xfdu/xsd-src/xfdu.xsd> – XFDU XML schema





XFDU again allows the possibility, taking advantage of XML accessor languages of transforming an AIP into a DIP, as with the SIP use of the sequencing and ordering of data and metadata files can be taken advantage of to automate splitting of file or re-grouping of files.

XFDUs built in support for **OAIS** Representation Information concepts and the availability of software development tool kits developed by NASA and ESA⁵⁰ make it a good candidate for a **CASPAR** packaging implementation, giving XFDU potential use in the role of Submission Information package, Archival information package or Dissemination Information package. XFDU is being used operationally by ESA as a specialized profile, SAFE (Secure Archive Format for Europe).

⁵⁰ <http://sindbad.gsfc.nasa.gov/xfdu/> – NASA XFDU JAVA toolkit; <http://www.gael.fr/xfdu/site/> – ESA Gael XFDU JAVA toolkit





7 DIGITAL RIGHTS MANAGEMENT AND ACCESS CONTROLS

7.1 OVERVIEW

In a general context, DRM deals with the definition and enforcement of rights to distribute and to use digital content, while AC deals with the enforcement of users' access to an Information Commons (IC) environment and its resources.

The objective of **CASPAR** is to re-use some existing and promising results in the state of arts. Thus, the outcomes of other projects have been analysed to investigate where they actually do provide adequate support to these facilities.

Different approaches have been adopted within the projects and it is an aim of **CASPAR** to see if they can work together. To this scope, the main positive and useful outcomes of some projects are summarized in the following sections.

7.2 SECURITY IN OTHER PROJECTS

Although they rely on different technologies and act at different levels, DRM and AC are closely tied. They are both related to security and protection of content, and they must cooperate in order to provide robust and at the same time fine-grained security policies. DRM licenses may be seen as an extension to AC policies within an archive; AC is enforced before DRM process begins.

For this reason, they have been analysed together, i.e., searching for cases where right and AC were both addressed and integrated in the same project.

In order to organize the process of evaluation, in particular to cover all relevant aspects of security, a proper set of metrics have been defined. In particular, the following security related aspects have been analysed:

- DRM policy creation
- DRM policy projection
- DRM security and cryptography
- AC technologies

Each of them is further divided into specific aspects, and is described in the next sections.

7.3 USEFUL REFERENCES

This section reports the most interesting outcomes of some other projects. It is a selection of significant references and it is organized in the four aspects defined above.

7.3.1 DRM policy creation

DRM can be organized in several stages. The three stages of DRM policy creation are:

- Recognition of rights
- Assertion of rights
- Expression of rights





7.3.1.1 Recognition of rights

Recognition of rights is the stage at which staff, employers and suppliers (i.e., publishers) all need to be aware of who the rights holders are (author, editor, compiler, institution, etc) and what uses they might be licensed for (play, copy, sell, etc).

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.1.1.1 Cedars

Still in the context of digital preservation, Cedars provided some guidelines that address even the preliminary phases of rights negotiation and licensing agreement. The legal framework considered in this project concerns general rights issues in the United Kingdom, with an emphasis on copyright and preservation licence issues.

7.3.1.1.2 DART

Privacy legal issues have been considered in DART, in particular those that arise in the scientific research area from information deposit into institutional repositories, and its subsequent storage, use and dissemination.

7.3.1.1.3 InterPARES

Concerning preservation rights InterPARES can be referred to, as it intends to promote legislation and policies to ensure long-term preservation and use of digital records. Guidelines and recommendations have been developed by the project with specific reference to the national and European legislations and to the international standards.

7.3.1.1.4 Variable Media Network

With regard of recognition of rights, an interesting contribution comes from Variable Media Network. In particular, this project has highlighted the necessity of deferring rights to source material, such as photographic negatives, video masters, Java source code, or the rights to modify or redistribute online works.

Another relevant outcome of Variable Media Network is recognition of novel issues that arise in particular in the digital art area and that concern the rights related to reproduce an artwork. In fact, an artist may wish or require that its work will be reproduced following his directives. These intentions of the artist may be part of the contract, thus need to be related to the rights to preserve the work. For instance, preservation information must capture the artists' desires about how to translate in future their work into new mediums.

7.3.1.1.5 CIDOC Conceptual Reference Model

CIDOC-CRM is an ontology. It may be extended/specialized to include some concepts of rights recognition.

7.3.1.2 Assertion of rights

Assertion of rights is provided by a legal framework in which people and organizations can assert their rights in a form that is defensible under law.

Several licence models are emerging that facilitate the assertion of rights by people that do not have the adequate skill for this task. Licence models allow adopting an existing licence such as those used for open source software, selecting a licence from a number of options available, such as Creative





Commons. The opposite consists in drawing up a licence specific to the institution, project or repository.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.1.2.1 BRICKS

In BRICKS support is given to create a syntactically correct digital license, or to import existing licenses, i.e., Creative Commons. These licenses are specified in the MPEG-21 Rights Expression Language (REL).

7.3.1.2.2 DART

DART investigated on how best to assist researchers in dealing with intellectual property issues during the research process, and addressed this issue developing a software to enable scientific researchers to attach standardized licenses that define attribution, commercialisation, derivative works and re-distribution rights. These licenses are specified in traditional human language.

7.3.1.2.3 CIDOC Conceptual Reference Model

CIDOC-CRM is an ontology. It may be extended/specialized to include some concepts of rights assertion.

7.3.1.3 *Expression of rights*

Expression of rights has traditionally involved only a copyright statement in a human readable form.

While this is still important, it is also essential to take account of machine-to-machine communication when considering DRM. This is achievable through Digital Rights Expression Languages (DREL), which allow the asserted rights to be expressed in a machine readable in addition to the human readable form.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.1.3.1 Archiving the Avant-Garde

Still related to metadata and rights management, a relevant requirement that emerges from the formal model developed in Archiving the Avant-Garde is the ability to describe subcomponents of a work, i.e., in the digital arts context. This allows a descriptive granularity that is useful in many ways, for instance (1) to include preservation metadata that may not apply to the whole work, (2) to manage digital rights as well as (3) to address AC with fine-grained policies. For instance, it allows specific designation of IPRs to specific images, and more in general, to different parts within a collaborative work.

As a conclusion, policy creation is generally not deeply addressed in digital management of rights, while it is policy projection that represents the main scope of DRM systems.

7.3.1.3.2 BRICKS

Concerning rights expression languages, in BRICKS DRM uses the MPEG-21 REL to define rights and conditions associated with the use and protection of digital content and services. MPEG-21 REL is probably the most widely adopted specification for DRM beside Open Digital Rights Language (ODRL). It was chosen due to its higher language expressivity and extensibility levels, in particular, any ODRL declaration can be mapped on MPEG-REL syntax.





7.3.1.3.3 Cedars

Concerning rights management and metadata, an interesting contribution is given by Cedars. The project proposed a significant expansion of the **OAIS** model and metadata. In fact, the Cedars metadata specification includes an extensive section on rights metadata, which records both the rights holders (and original copyright statements and warnings) and the allowed actions (and the actors permitted to perform these). Intellectual Property Rights (IPRs) are seen as part of Provenance Information, thus extend the PDI metadata of the **OAIS** model.

7.3.1.3.4 CIDOC Conceptual Reference Model

CIDOC-CRM is an ontology. Machine processable licenses are out of its scope. It can be extended/specialized to include licensing metadata.

7.3.2 DRM policy projection

The three stages in DRM policy projection are:

- Dissemination of rights
- Exposure of rights
- Enforcement of rights

7.3.2.1 Dissemination of rights

Dissemination of rights ensures that wherever a resource is described its rights are also described. This is the first stage in projecting the DRM policy. It requires that when resource hubs gather metadata they also gather rights metadata. Typically, a rights element will contain a rights management statement for the resource, or reference a service providing such information. Thus, rights dissemination consists of providing all the digital rights metadata in conjunction to content.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.2.1.1 BRICKS and Cedars

Dissemination is closely tied to expression of rights. Thus, Cedars and BRICKS are still the two main references for dealing with issues that concern rights metadata.

7.3.2.2 Exposure of rights

Exposure of rights is the stage at which a user will see the rights information associated with a resource.

Exposure of rights should be evident when searching for resources, i.e., it should be clear, usable and transparent. If there are differences between the permitted uses for different objects then these should be easily apparent without detailed scrutiny of licence conditions. To this scope, terms or symbols that have a clear and unambiguous meaning should become familiar so that understanding rights information can become effective.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.2.2.1 BRICKS

In BRICKS rights information is embedded in digital content storage, in particular by means of XML licenses specified in a formal declarative REL, namely MPEG-21 REL.





7.3.2.2.2 Cedars

In Cedars the rights management metadata is designed to provide a qualified human being with enough information to decide whether a user should be given access to a resource or not.

7.3.2.2.3 DART

In DART the rights are exposed through traditional licences expressed in human language. The consumer just reads and accepts term and conditions of usage.

7.3.2.2.4 CIDOC Conceptual Reference Model

CIDOC-CRM is an ontology. Usage rights can be included (see Recognition and Assertion of Rights). Mechanisms for exposure of rights are out of its scope.

7.3.2.3 *Enforcement of rights*

Enforcement of rights includes both protective measures to ensure that rights are not infringed and steps to be taken when infringements are detected.

The most basic protective measures ensure that access to resources is granted only to people who have acknowledged that they have accepted the licence conditions under which the resources are made available. At this basic level, a system of authentication and authorization is required.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.2.3.1 BRICKS and Cedars

Mandatory enforcement of digital rights is best addressed in BRICKS, where the DRM Layer within the overall system architecture implements this service at several levels, as it cooperated with the Security Management Layer to block every user operation that can violate rights defined in a machine-readable format. Cedars provides support for rights enforcement, too.

7.3.3 DRM security and cryptography

The following aspects of DRM security and cryptography have been taken into account:

- Digital identity management
- Integrity
- Copy control and protection
- File formats

7.3.3.1 *Digital identity management*

Identity management is part of authentication, namely the process that asserts that an individual or an entity is who it claims to be. Digital identity may for instance be represented by a user name, optionally secured with password. In the context of DRM identity is proven through certificates. Certificates contain user public key and owner identity information. Digitally signed certificates are provided by certification authorities (CA), and use the digital signature to guarantee the identity of the authority that has issued them.





(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.3.1.1 BRICKS and Cedars

In DRM, an adequate digital identity management, in particular the implementation of such feature has been found in BRICKS and Cedars. Here digitally signed certificates are allocated to the users when they log in. Then, identity information may, for instance, be used when applying the watermarking technique that allows embedding identity information of both IPR holders and consumers.

7.3.3.2 Integrity

Integrity control can be addressed at several levels, for example through checksum, digital signatures, or digital watermarks. It is worthwhile noting that, in the preservation context; integrity does not only mean integrity of bit streams, but in particular of information and semantics of a digital content.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.3.2.1 Chronopolis

Integrity metadata in Chronopolis also contain information on audit trails and AC.

7.3.3.2.2 DILIGENT

Several projects recognized the necessity of safeguarding digital content integrity. Among these, with focus on digital preservation, DILIGENT supports integrity control using watermarking and signing.

7.3.3.2.3 InterPARES

In the case of InterPARES the integrity is not intended only as a tool to ensure the verification of the unchanged bitstream within the repository but also as a series of procedures and documentation of the processes involved in the preservation. A specific analysis of this aspect will be specifically developed in a position paper on authenticity.

InterPARES uses digital signatures and checksums to verify that between transformation events, the digital entity has remained unchanged, and the mechanisms used to accession records can be re-applied to validate the integrity of the digital entities between transformative migrations. InterPARES integrity is implemented via appropriate metadata.

7.3.3.2.4 Provenance

A very particular approach is taken in Provenance, which suggests a novel solution for annotation and tracking the history, namely the provenance, of digital content. In short, the challenge of tracking digital data history relies on provenance assertions or “p-assertions” that are produced and recorded in a “provenance store”. By enabling actors to make execution-related assertions, or p-assertions, which ensure that necessary and sufficient forms of process documentation are captured, the system allows to give a complete account of any data item’s provenance. For example, the p-assertion model allows to document various aspects of execution, and thus provide descriptions of those parts of an execution that relate to, or impact upon, a given data item. This allows a user to determine the data item’s relationships to other data items and processes, such as its dependencies or causal effect, which in fact may result useful in dealing with integrity control of digital objects subjected to preservation transformations. The documentation related to the provenance is a relevant component in the process to ensure the integrity, to provide evidence of the preservation tools and information for maintaining the history of the resources over time.





7.3.3.3 Copy control and protection

Copy control and protection is part of rights enforcement, and should still be provided externally to the platform, once a digital content has been physically delivered to the final user.

The most important technological measure for copy control and protection is the watermarking technique, which comprises several facilities. Watermarking allows, for instance, embedding identity information of both IPR holders and consumers or referencing the license acquired to obtain a content, it allows, together with a special type of encryption, to alter the media quality in a controlled manner, i.e., without its full loss, in order to provide proper quality only for the granted usage or to granted users. Watermarking technique in short supports content authenticity, integrity and control on copy usage through the embedment of control information inside of the digital content.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.3.3.1 BRICKS and DILIGENT

Watermarking for IPR protection is implemented in BRICKS and in DILIGENT projects

7.3.3.4 File formats

Concerning DRM, it is mainly copy control and protection that may impact on the file formats. In fact, protection measures that are applied outside the environment must be embedded in the digital content.

7.3.3.4.1 BRICKS

In BRICKS watermarking is applied only on image files.

7.3.3.4.2 Planets

In the context of preservation, Planets states that it will use a particular file format to store digital assets, namely the file format of IBM's UVC solution. The advantage of this restriction is that longevity guaranty of the assets is addressed. Although DRM is performed at a different level with respect to preservation, restriction on file formats that come from preservation necessities should be considered.

7.3.3.4.3 Preserv

A file format related utility has been used in Preserv, namely the freely available PRONOM software for identification and verification of file formats. It is a web-enabled database of information on file formats and their technical dependencies, including hardware, software and operating systems.

7.3.4 Access control (AC) technologies

The considered aspects in AC are:

- Granularity
- Identity management
- Interoperability
- Level of security





7.3.4.1 Granularity

Access policies are in most cases enforced through a role based AC. However, there is a relevant improvement in the granularity of AC, if interoperability with digital rights mechanisms is provided. As a consequence, a machine-readable format for rights specification is basic assumption.

(A) PROJECTS AND OTHER MAJOR INITIATIVES

7.3.4.1.1 BRICKS

BRICKS uses MPEG-21 REL and Cedars proposed an extension of OAIS metadata for rights management. A fine-grained AC has been found in BRICKS, where the Security Layer collaborates with the DRM Layer to ensure adequate enforcement of rights protected content (and services). In fact, authorization is handled at two different levels: a static one that defines basic policies for accessing services and content, i.e., by checking the user-provided signed certificate, and gathering which groups the user belongs to along with roles associated to them, and a dynamic one that overrides the static policies if particular conditions are met (i.e., a fee is paid for accessing the service). Thus, AC is strictly linked to Accounting and IPR modules.

7.3.4.1.2 Cedars

Integration of right policies in AC has been addressed in Cedars, too. An archivist is allowed to specify the permitted users of the digital object, for example, archive staff or library users or both, and furthermore, it is allowed to decide what a library or archive user can do with a digital object. From the implementation perspective, a digital certificate is allocated to each user. Where an AIP is placed in restricted access, a reference to an access contract is placed in its gateway file. When a user makes a request to download one of the DIPs for this resource, if the user has a Cedars digital certificate that is listed in the access contract then the request is granted, otherwise the download is prevented. Whenever a new user is given permission to access resources, their digital certificate number is added to the relevant access contracts. A special tool is used to specify an access contract.

7.3.4.1.3 Provenance

It is worthwhile mentioning Provenance AC solution that considers provenance information, i.e., p-assertions, to block unauthorized operation. In fact, control is not only based on the authorizations specified in the authorization policy (i.e., through certification) and the (role-based) internal representation of the user, but may additionally be dependent on information contained within the data item that the request is related to. Such a condition is specified accordingly in the authorization policy. For example, a read operation associated with an internal representation (of the user) on a given p-assertion might be permitted only if the p-assertion contained relevant information pertaining to that internal representation. In this case, the p-assertion in question would have to be retrieved first and assessed accordingly by the authorization engine before a final decision can be made on granting or denying the request.

7.3.4.1.4 CIDOC Conceptual Reference Model

CIDOC-CRM is an ontology. It could be extended to include some administrative directives (e.g. access control specifications) with the desired granularity.

7.3.4.2 Identity management

Identity may be managed for instance through a username to log in an environment. However, such credential is rather weak, and in particular, it loses validity once this information is referenced outside the environment. A stronger identity credential is provided through digital certificates.





Depending on the certification authority that signs the certificate, identity may become more or less trusted and reliable, and in particular, such identity credential is system-aware.

(A) *PROJECTS AND OTHER MAJOR INITIATIVES*

7.3.4.2.1 BRICKS and Cedars

Identity management based on digital certificates are used, where DRM is part of the AC.

7.3.4.2.2 Chronopolis, DART, DILIGENT, InterPARES and Provenance

Identity management based on digital certificates are also used in other projects that rely on the GSI. In particular InterPARES has developed *benchmark requirements* to support a presumption of authenticity which include attributes/elements specifically related to the identity of the resources acquired by the preserver.

7.3.4.3 Interoperability

Interoperability is considered here with respect to different AC solutions. It is particularly interesting to make policies interoperable, when they need to be managed independently, for instance, to integrate the policies to access the environment, to perform operations once a user has logged in, and finally, to control operations on right-protected archived items.

(A) *PROJECTS AND OTHER MAJOR INITIATIVES*

7.3.4.3.1 BRICKS and Cedars

In BRICKS, two different authentication mechanisms interoperate, as users log in the environment providing username and optionally a signed certificate that is used to implement rights related AC. Similar feature is supported in Cedars, where digital certificates are allocated for each user, and access policies can be easily extended intervening on the management of the access contracts that are allocated to the archive resources.

7.3.4.3.2 DART

In DART authorization relies on a role based AC associated with the GSI, namely a Public Key Infrastructure (PKI) for identification using X.509 certificates and Globus gatekeeper facilities.

7.3.4.4 Level of security

(A) *PROJECTS AND OTHER MAJOR INITIATIVES*

7.3.4.4.1 BRICKS

BRICKS uses watermarking, signing and encryption applied to single digital content resources.

Integration of DRM in access policies is realised in BRICKS.

7.3.4.4.2 Cedars

Integration of DRM in access policies is realised Cedars.

7.3.4.4.3 Chronopolis





Chronopolis may have useful experience with GSI.

7.3.4.4.4 DART

DART may have useful experience with GSI.

7.3.4.4.5 DILIGENT

DILIGENT may have useful experience with GSI.

DILIGENT uses watermarking, signing and encryption applied to single digital content resources.

7.3.4.4.6 InterPARES

InterPARES may have useful experience with GSI.

7.3.4.4.7 Provenance

Provenance may have useful experience with GSI.

Integration of DRM in access policies is realised, to some extent, in Provenance.





8 PROOF OF PRESERVATION EFFECTIVENESS

(A) PROJECTS AND OTHER MAJOR INITIATIVES

8.1 BRICKS

The BRICKS Framework, distributed under the LGPL license, has been designed from the beginning to be open, extensible and based on standards.

All BRICKS services adopt the most relevant and accepted standards for representing content, metadata, administrative information, schemas, access control policies, digital licenses, annotations, etc.

Standard web service technologies are used for communication with BRICKS services, allowing for interoperability with other implementations, and guaranteeing an easy upgrade path for future extension. The advantages of a standard-based approach are evident, as any part of the system may be easily reused and is fully documented.

The modular BRICKS application may allow the implementation of vertical preservation services, and be adapted to suit the specific preservation requirements, either by incorporating existing components such as registry, etc, or by replacing the existing services implementations with preservation aware ones, thus affecting only limited parts of the system.

In conclusion, BRICKS may be considered as a promising open framework to serve as a basis for a preservation system, as it offers the necessary infrastructure for the development of targeted preservation services, and partially provides some of the feature needed for an OAIS implementation. Missing features can be easily added with the support of the BRICKS Communities, thanks also to the liberal licensing terms.

8.2 DILIGENT

DILIGENT builds on an open services architecture, which means it uses services and components developed with open standards, so it could be easily reusable in case of technological changes.

The framework relies on a set of standards and technologies (i.e., WSRF framework, some WS-* standards, WS standards - XML, SOAP, and Web Services Description Language (WSDL)), well-defined classes and packages have been identified in the system design and well known software engineering techniques have been used. Therefore the system is capable to cope with technological changes simply because of its design.

For instance, each service has a class taking care to implement the communication mechanism. If this mechanism should change it is enough to modify/replace this class in order to ensure the system operation.

Concerning the information, the services have been designed to deal with generic objects represented in XML, and part of their code (appropriately isolated and factored) takes care to transform those objects in internal data structure. Technologies changes would affect only part of the services and thus can be easily tackled.

DILIGENT can be seen as general-purpose architecture, potentially able to support any type of user requests, independently of the specific contents and services. The infrastructure is designed to deal with dynamic community whose requirements change over the time. The test-bed scenarios provide user communities with DLs based on the aggregation of resources explicitly required by the users themselves. For instance, DILIGENT supports the addition of new resources (either information objects or services/applications) to the resources pool made available to the users.

8.3 FEDORA





A Fedora repository can, surrounded by the proper preservation policies, tools, and Fedora services, serve as the basis of a trustworthy preservation system. Fedora cannot serve as the entire preservation system, but only as a preservation application, which is just a portion of the entire system. Without the appropriate people, infrastructure, policies, and procedures, even the best preservation application cannot ensure preservation.

The Fedora core provides a promising basis for a preservation system. Its agnostic view towards file formats and object types enables it to manage essentially any type of file. It has the ability to manage objects with complex — including hierarchical — relationships with its use of RDF or Metadata Encoding and Transmission Standard (METS) metadata. It can manage multiple bit streams for a single object, which can enable archivists to track and store the original bit stream of an ingested record and the bit streams of subsequent transformations. It has versioning and persistent identifier capabilities for all content objects, metadata, and disseminators. With eXtensible Access Control Markup Language (XACML), an institution can articulate policies to help manage access to records. Fedora is a transparent system and Fedora objects are articulated in XML (usually FOXML or METS), making it feasible to migrate records out of Fedora.

8.4 SAFE

The adoption of a common format like SAFE gives major benefits for a cost-effective long-term preservation and exploitation of these data for several reasons:

- SAFE is mainly devoted to the long-term preservation of data, as its full compliance with the CCSDS/ISO OAIS RM and XFDU standards demonstrate.
- SAFE permits an easier and more effective migration of data to other future standards.
- SAFE permits and easier reformatting to other formats, including end-user formats for products distribution.
- SAFE will ease the maintenance of the SW that use the data, even historical, thus decreasing the chances of obsolescence and improving their usability.
- SAFE being self describing, it greatly facilitates the format maintenance.
- SAFE format documents are built automatically from the XML-Schemas, so the maintenance of the SAFE format is more robust with respect to “old-fashion” formats.

Software (SW) that access SAFE-formatted datasets will benefit from the usage of the same SAFE XML Schemas for reading and writing the datasets. This greatly improves the overall cycle of creation/ingestion/quality control/transformation/distribution of the datasets.

The SAFE format provides a structure for long-term archiving of the ESA’s historical and future datasets. The main operational activity presently carried out is the archiving of the datasets from the present media into the automated archive, in their original format. The HARM project will deliver the SW to perform the selection of the overlapping datasets, their conversion into SAFE and their stitching whenever an overlap reduction is required. By doing so ESA expects to greatly reduce the effort required maintaining its data holdings and improving the long-term preservation of the data.

The adoption of SAFE for the ESA’s historical datasets in order to safeguard them from the possible loss is just the first step of the ESA’s strategy for the long-term preservation of its EO missions’ data archive. The future adoption of SAFE as archiving format since the design phase of the future missions will enhance even more the capability of ESA to preserve these important data, thus guaranteeing their accessibility and usability by the future generations. With this perspective, ESA welcomes and supports activities outside of its organization for the adoption of SAFE as standard exchange archive data format and/or exchange format.





ESA is in course of establishing a SAFE dedicated web site where the documents, APIs and other information will be made available to external users.⁵¹

⁵¹ At the time of writing this paper, the SAFE web site is under construction and reachable at the following URL: <http://earth.esa.int/SAFE/>.





9 SUMMARY

It is clear that there are a great number of useful ideas and techniques which can be taken from other projects, however it is also clear that there is still much which **CASPAR** needs to do in order to contribute to a complete end to end preservation system.

Some of the projects and initiatives surveyed are partly **OAIS**-compliant. They might have informal processes equivalent to ingest, or implement specific parts of the **OAIS** standard (e.g. PAIMAS). Various workflow mechanisms have been implemented to assist with, for example, the production of Archival Information Packages. Some areas are better covered than others: not surprisingly, data storage and data management are well represented. In some domains such as digital arts, preservation planning is being considered in response to the pressing needs for preservation. In the science domain, there are languages and formats for describing scientific data.

Finally, we note that this deliverable is a snapshot of a living document which will be updated and made available on the **CASPAR** public web site. In part this will reflect the lessons learned through the implementation phase as part of the iterative development of the project deliverables. It will also reflect contributions from outside the project which provide answers to the questionnaire on which this document is based. As such this document, plus the collection of Annexes, should provide a rich resource for other projects, especially those which wish to adhere to an **OAIS**-based approach to preservation.

