



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

CASPAR DRAFT TESTBED IMPLEMENTATION PLAN



Document identifier:	CASPAR-4105-RP-0101-1_2
Submission Date:	15-01-2009
Due date:	15-01-2009
Work package:	4100
Partners:	All Partners
WP Lead Partner:	STFC
Document status	FINAL

Abstract: This document provides the implementation plan for the CASPAR testbed.



Delivery Type Report
Author(s) CASPAR Consortium

Approval David Giaretta

Summary

Keyword List

Availability PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0_0	30 June 2008	Material gathered on Wiki	Whole consortium
0_1	10 Jan 2009	Word version	David Giaretta
1_0	13 Jan 2009	Almost final version	David Giaretta
1_1	15 Jan 2009	Revisions from data holders	Sergio Albani, Fulvio Marelli, Mariella Guercio, Esther Conway
1_2	15 Jan 2009	Final version	David Giaretta





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

Project Co-ordinator: David Giarretta

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





CONTENT

1	INTRODUCTION	6
2	METHODOLOGY	7
2.1	GENERIC CRITERIA AND METHOD TO ORGANISE AND TO EVALUATE THE TESTBEDS	7
2.1.1	<i>Method</i>	7
2.1.2	<i>Criteria</i>	8
2.1.3	<i>Test definitions</i>	8
2.1.4	<i>About authenticity</i>	16
3	CULTURAL TESTBED – UNESCO	17
3.1	PURPOSE OF THIS SECTION	17
3.2	CONTEXT	17
3.3	CASPAR AND DIGITAL PRESERVATION	18
3.3.1	<i>Process details</i>	19
3.3.2	<i>Concrete example for the Testbed</i>	19
3.3.2.1	<i>The process flow</i>	20
3.4	THE CULTURAL TESTBED SHOWCASE APPLICATION	23
3.5	VILLA LIVIA DATASET OVERVIEW	28
3.6	ESRI ASCII GRID FILE: <i>DEM_LOD3_LIVIA.GRD</i>	29
3.6.1	<i>Structural and Semantic RepInfo for ESRI GRID file format</i>	29
3.6.2	<i>Preservation Description Information for an ESRI GRID file AIP</i>	31
3.6.3	<i>Descriptive Information</i>	31
3.6.4	<i>Complete AIP for ESRI GRID file</i>	31
3.7	ESRI SHAPE FILE: <i>VINCOLI LIVIA.GRD</i>	32
3.7.1	<i>Structural and Semantic RepInfo for ESRI Shape file format</i>	32
3.7.2	<i>Preservation Description Information for an ESRI Shape file AIP</i>	33
3.7.3	<i>Descriptive Information</i>	33
3.7.4	<i>Complete AIP for ESRI Shape file</i>	33
3.8	RELATED DOCUMENTATION	33
3.9	OTHER MISC DATA WITH A BRIEF DESCRIPTION.....	33
3.9.1	<i>UNESCO Glossary</i>	34
3.9.2	<i>UNESCO References</i>	34
4	CONTEMPORARY PERFORMING ARTS TESTBEDS.....	36
4.1	IRCAM TESTBED	36
4.1.1	<i>Test cases</i>	36
4.1.2	<i>Test scenarios</i>	36
4.1.3	<i>Definition and setup of testbed infrastructure</i>	41
4.1.4	<i>- Training, production of documentation</i>	42
4.1.5	<i>- Definition of scenarios and validation support</i>	42
4.1.6	<i>- Tests and validation</i>	42
4.2	UNIV LEEDS TESTBED	44
4.2.1	<i>Test cases</i>	44
4.2.1.1	<i>Demo data and Representation Information</i>	44
4.2.1.2	<i>DC Profiles</i>	44
4.2.2	<i>Scenarios</i>	44
4.2.2.1	<i>Scenario 1: Creation of a new Information Package</i>	44
4.2.2.2	<i>Scenario 2: Retrieve an Information Package</i>	45
4.2.2.3	<i>Scenario 3: Updating an Information Package</i>	45
4.2.2.3.1	<i>Strategy 1 (update the old package):</i>	45
4.2.2.3.2	<i>Strategy 2 (create a new package and relate it through comments with the old one):</i>	46
4.2.3	<i>Testbed Phases</i>	46
4.2.3.1	<i>Phase 1: Definition and setup of testbed infrastructure</i>	47
4.2.3.2	<i>Phase 2: Scenario Definition</i>	48
4.2.3.3	<i>Phase 3: Key Components Integration</i>	49
4.2.3.4	<i>Phase 4: Training</i>	49
4.2.3.5	<i>Phase 5: Tests</i>	50
4.3	CIANT TESTBED	51





4.4	INA-GRM TESTBED	52
5	SCIENCE TESTBEDS.....	54
5.1	STFC TEST BEDS	54
5.1.1	<i>OAIS Preservation information flow diagrams</i>	55
5.1.2	<i>Information Objects</i>	55
5.1.3	<i>Supply Relationship</i>	56
5.1.4	<i>Supply Process</i>	56
5.1.5	<i>Packaging relationship</i>	56
5.1.5.1	Information object dependency relationships	57
5.1.6	<i>Preservation strategies</i>	57
5.1.6.1	In response to a supply impediment	57
5.1.6.2	In response to an identified information preservation risk	57
5.1.6.3	As a secondary response to a preservation strategy	57
5.1.7	<i>The Implementation Plans</i>	58
5.1.7.1	Implementation plan for Scenario1 MST-Simple.....	58
5.1.7.1.1	Preservation Information Flow for Scenario1 MST-Simple	59
5.1.7.1.2	Implementation points based on strategies for scenario1	59
5.1.7.2	Implementation plan for Scenario2 MST-Complex	60
5.1.7.2.1	Preservation Information Flow for Scenario2 MST-Complex	61
5.1.7.2.2	Implementation points based on strategies for scenario1	61
5.1.8	<i>Ionosonde data and the WDC</i>	61
5.1.8.1	Implementation plan for Scenario3 Ionosonde-Simple	62
5.1.8.1.1	Preservation Information Flow for Scenario3 Ionosonde-Simple	63
5.1.8.1.2	Implementation points based on strategies for scenario3	63
5.1.8.2	Implementation plan for Scenario4 Ionosonde-Complex	63
5.1.8.2.1	Preservation Information Flow for Scenario4 Ionosonde-Complex	65
5.1.8.2.2	Implementation points based on strategies for scenario4	65
5.2	ESA.....	66
5.2.1	<i>Abstract</i>	66
5.2.2	<i>Relation with Deliverable 4101</i>	66
5.2.3	<i>Testbed Description</i>	66
5.2.3.1	Background and purpose	66
5.2.3.2	The Testbed Phases	67
5.2.4	<i>Testbed CASPAR components</i>	67
5.2.5	<i>Testbed development time schedule</i>	69
5.2.6	<i>Gome dataset and designated communities</i>	70
5.2.6.1.1	Gome Dataset Definition	70
5.2.7	<i>ESA Scientific Testbed Scenarios</i>	71
5.2.7.1	Data Ingestion	71
5.2.7.2	Data Search and Retrieval	72
5.2.7.3	Level 1C data creation.....	72
5.2.7.4	Software change	73
5.2.7.5	New Data browsing	73





1 INTRODUCTION

The purpose of the CASPAR testbed is to provide evidence that the CASPAR approach is doing something useful for digital preservation in several different domains in several different organisations, and forms part of the set of metrics described in the CASPAR Description of Work [DoW]. The testbeds also provide an indication of how CASPAR might be used in the longer term within organisations.

D4101 USER REQUIREMENTS AND SCENARIO SPECIFICATIONS provided a wide collection of scientific, artistic and cultural data together with scenarios capturing threats to their continued usability and access. This document follows on from D4101 by providing, for a selection of the D4101 scenarios, details of how these are to be carried out and, very importantly, how we will validate the results.

This document first provides a summary of the methodology adopted then a summary of the initial set of tests. Following is a collection of contributions from various authors, providing sections devoted to each of the disciplines and organisations, giving more details of the background, the test set-up and evaluation criteria for the initial sequences of tests plus a tranche of others; further tests will be added as time permits.

It must be recognized that there will be a number of iterations in the process. We are testing a dynamic, developing, open system. For each dataset there will be a number of variants which must be investigated, for example a number of ways of capturing Representation Information are usually possible. It may be that some new techniques have to be introduced. Thus the method to evaluate the success or the compliance of the testbeds will be based on an **iterative process of tests and feedback**.

This iterative process introduces the risk that any piece of evidence we produce will be less convincing than we would wish. However it is hoped that the accumulation of evidence will provide adequate reassurance.



2 METHODOLOGY

2.1 GENERIC CRITERIA AND METHOD TO ORGANISE AND TO EVALUATE THE TESTBEDS

The goal of this section is to describe the generic criteria and method for the evaluation of the testbeds.

2.1.1 Method

For the most part only Designated Community members can really evaluate the preservation results by access to and manipulation of the data, therefore individuals will have to be identified to provide this level of validation. However many aspects can be judged by data manager colleagues of the CASPAR team. It is expected that testbed specific criteria will have to be defined to supplement a number of generic ones. In any case it should be relevant to consider the validation processes from an auditing point of view.

Scenarios defined in the D4101 have to be instanced within the testbed by going from the easiest to the most complex ones. As noted in the CASPAR Conceptual Model, we can consider that for each testbed the specific scenarios may be divided into three generic kinds:

A the *environment* may change

this includes all the general environment that could alter or disappear

- artistic testbed: e.g. the musical environment and network resources could change
- cultural testbed: e.g. national boundaries could change
- scientific testbed: e.g. open source software may no longer be supported by the community

B the *hardware and software environment* may change

this includes all the elements of hardware and software supporting data use

- artistic testbed: e.g. microphone, loud speakers, sensors, NEXT station, Protocols, MAX/MSP.
- cultural testbed: e.g. the classification papers have disappeared, pdf format is no longer an ISO standard format, the electronic system for ingesting and/or managing the resources is no longer supported
- scientific testbed: GOME Data Processor Extraction Software out of rights licensing

C the *knowledgebase of the Designated Community* may change

because we don't have time to wait for a natural evolution of the DC, a simulation may be radically a change of DC (furthermore a change of the foreseen usages, so that a



change of the expectations about the archives information). In general the approach will be for current members of the Designated Community to develop sets of questions which probe whether or not someone understands that dataset, and then check that adequate additional Representation Information can be created to help people with different knowledge backgrounds. This is without doubt going to be an extremely difficult exercise.

Closely associated with the DC changes we need to recognize related changes in common or community knowledge: the data *production team implicit knowledge* will be lost.

--> this concerns all the information that has not been explicitly formalized during the production of the data

- artistic testbed: e.g. interactions between dance performers...
- cultural testbed: e.g. non documented reasons of a classification needed after the death of the decision maker
- scientific testbed: e.g. exceptional use of a particular radiance sensor in a place has not been made explicit

2.1.2 Criteria

Here are the main generic criteria to evaluate the success of the testbeds:

1. at least one testcase of data is necessary for each of the preservation scenarios
2. for each testcase/scenario a strategy of preservation must be determined and applied
3. for each strategy the link with CASPAR components must be thought through, including toolkit aspects such as authenticity
4. for each strategy must include use and feedback by member of the appropriate DC
5. All the results will contribute to the validation report (D4103)

N.B.: The running of a test may need DC members to access and manipulate the archives in order to satisfy the predetermined usages, so that a user interface of some kind may be implemented. Specific metrics are expected for each testbed in order to help to evaluate how well the validation criteria will have been met.

2.1.3 Test definitions

A number of the early tests are summarized in the following table.

.



Scenario ref	Person responsible	Type of change	Scenario description	Setup e.g. dataset specification	Needed from CASPAR	Process	Final state	Validation test	Success indicator
Reference to section/scenario in D4101	Named individual	Hardware/ Software/ Environment / Designated Community	Description of scenario, based on D4101 text	Specific dataset or other setup conditions	What CASPAR system would need to produce e.g. Representation on Information for specific dataset	How this is created and applied	What we should have at the end of the test	How we check that the information is preserved	How we will know that the preservation has been successful.
D4101 – chapter 4.1.4.1: Scenario 3 & 4	Raffaella Ciavarella	Hardware Software	* An archivist ingests a new WORK into MustiCASPAR² ; * A registered expert sends notification of loss of availability for a COMPONENT; * A registered expert	080429_scenario_data see here	- Ability to store, preserve RepInfo for spat component - Ability to generate and preserve an authenticity protocol for spat component - Ability to manage notifications	see here	A new version of « Avis de Tempête » (spat component replaced)	- Availability of information about obsolescence of spat - Ability to generate or localize a new version of spat - Validation of new version of spat by Audio engineer by	- Ability to reperform the work





			receives an asynchronous notification of loss of availability for a COMPONENT; * A registered expert searches for equivalent COMPONENTS; * An archivist generates a new version of the COMPONENT for an open-source RT engine (PureData?) from the stored virtual version and makes it available (applied		(sent by POM) about obsolescence of spat component			executing an Authenticity protocol	
--	--	--	---	--	--	--	--	------------------------------------	--





			strategy #7); * An archivist verify the authenticity protocol of the old COMPONENT, for all the existing WORKS containing it; * An archivist ingests the new version of the unavailable COMPONENT.					
D4101 3.4.4.1	SergioAlbani / FulvioMarelli	Changes in hardware and software	Change of compilers/libraries/drivers affecting ability to run the GOME Data Processor L1b->L1c - Implementation of Scenario	see DATA section below	Need to preserve the ability to generate L1c data starting from L1b data.	see details below		Ability to process L1b data to produce





			1 regardin all the Sip/AIP creation, ingestion and data search and retrieval based on different user profiles. See the implementation plan for details.						
MST scenarios	Esther Conway /Matt Dunckley (STFC)	Change in RepInfo	The MST NetCDF data files reference CF (climate forecast) standard names defined in the CF XML document, this document is defined by the climate community and new versions are	(NetCDF formatted) MST cartesian version 3 wind profiling data	Preserve the meaning and understanding of the Climate Forecast standard names	MST scenarios	A new version of CF-standard names file stored within the registry, references to the new repInfo from the MST AIPs. A notification detailing the change must be sent to all MST	Availability to access of new repInfo	Notification received from POM, access to new repInfo demonstrated





			released from time to time. A subset of these CF standard names in the CF XML document are specifically referenced by the MST data files, for example the 'tropopause_altitude' standard name. When a new version of the CF standard names document is released with a change to a CF standard name or its associated description used in the				stakeholders registered as part of the MST community.		
--	--	--	---	--	--	--	---	--	--





			MST data, the associated CF RepInfo needs to change to reflect this update. The CF standard names XML file is SEMANTIC repInfo, it provides meaning to the climate file parameters found within the MST data file.					
Cultural Testbed scenario 3	Davide Palmisano Patrick Min	Change of File Formats	A set of ESRI Shape files that cannot be visualized due to the unavailability of a suitable software	see below	Preserve the meaning and understanding of the ESRI Shape files			according to the information retrievable by the RepInfos attached to the ESRI Shape File, we have to demonstrate





								that a member of the DC can understand it	
--	--	--	--	--	--	--	--	---	--





2.1.4 About authenticity

Testbeds must provide occasions to test CASPAR components. And this is true especially for the Authenticity Management tool since the viability or the relevance of the system depends mainly on the evidence of archives/data authenticity. To improve its own protocol, authenticity annotation process must be integrated and tested as soon as possible along the production, the archiving, the access and the manipulation of the data.



3 CULTURAL TESTBED – UNESCO

by Davide Palmisano (ASMX), Maria Rosa Cárdenas, Mario Hernández (UNESCO)
with acknowledgements for
Roberto Scopigno and Marco Callieri from CNR-ISTI Visual Computing Lab.

3.1 PURPOSE OF THIS SECTION

The purpose of this document is to identify and describe a case study relating to the class of processes collectively entitled ‘Virtual Heritage reconstruction processes (VHRP)’ This has been done in order to have a concrete complete example so that we could as a team select interesting steps for ‘digital data preservation’ and to include this for the CASPAR testbed exactly as it was originally planned, except that now handling a concrete example.

Following a description of a concrete instance of VHRP, an implementation plan for a showcase application, named EWE[1], is provided.

3.2 CONTEXT

Processes collected under the VHRP class are generally aimed at the virtual-digital reproduction of cultural heritage.

According to the UNESCO Convention text [2], cultural heritage are:

- *monuments: architectural works, works of monumental sculpture and painting, elements or structures of an archaeological nature, inscriptions, cave dwellings and combinations of features, which are of outstanding universal value from the point of view of history, art or science;*
- *groups of buildings: groups of separate or connected buildings which, because of their architecture, their homogeneity or their place in the landscape, are of outstanding universal value from the point of view of history, art or science;*
- *sites: works of man or the combined works of nature and man, and areas including archaeological sites which are of outstanding universal value from the historical, aesthetic, ethnological or anthropological point of view.*

Hence, the definition of a generic VHRP will rely upon those processes that use technologies in the fields of 3D computer graphics and virtual reality to achieve the digital reconstruction of monuments and sites. One of the best definitions is that given by Maurizio Forte, one of the creators of the Virtual Museum of Ancient Via Flaminia [3,4], who states that:

a virtual heritage is a digital dynamic information that is derived from a physical site or intangible activities heritage, whatever is not just monument territorial landscape.

VHRP identifies a broad range of different processes that can vary from technological to methodological points-of-view such as: photogrammetry, time-to-flight laser scanning, triangulation scanning, 3D modelling and GIS integration.

The 3D laser scanner is an active scanner (hardware) that uses laser light to probe the environment. 3D triangulation-based laser scanners shine a laser on the subject and use



a camera to locate the laser dot. Depending on how far away the laser strikes a surface, the laser dot can appear at different places in the camera's field of view.

This technique is called triangulation because the laser dot, the camera and the laser emitter form a triangle. The length of one side of the triangle, the distance between the camera and the laser emitter is known. The angle of the laser emitter corner is also known.

The angle of the camera corner can be determined by looking at the location of the laser dot in the camera's field of view.

These three pieces of information fully determine the shape and size of the triangle and give the location of the laser dot corner of the triangle. In most cases, a laser stripe, instead of a single laser dot, is swept across the object to speed up the acquisition process.

3D triangulation laser scanning is the first task in the reconstruction process. Starting from an existent and physical object, the scanner produces the 3D data model of the object. Such data will be subjected to further elaboration in the next steps, producing new data and metadata, leading to a digital artefact such as a movie or a navigable 3D environment. The rest of this document will explore all the sub-tasks.

The follow points were considered in the choice of this specific process:

- the complex process makes possible positive assumptions regarding the significance of the specific instance. It allows identification of certain specific uses cases that can be easily integrated within the CASPAR Cultural Testbed scenario[5];
- data and digital items produced during the overall process have already been subjected to study within the framework of the CASPAR Cultural Testbed activities[6]
- this specific type of process is an emerging field in the cultural heritage domain and is starting to be heavily used. However associated digital-data-preservation is for the moment not a main consideration. Therefore it requires heavily investigated a good framework for the digital preservation of all the digital data produced, the hardware and software used and the process and knowledge that need to be followed.
- Its value has also been recognized within the CASPAR Project in relation to the Bamiyan Buddhas [7]. Since the destruction of this cultural heritage the digital data describing them represents the only knowledge available to human beings.

3.3 CASPAR AND DIGITAL PRESERVATION

We have learnt while working with CASPAR, that digital preservation implies preserving the associated 'bit strings' but since this is not sufficient in order to ensure full preservation, the concept needs also to preserve the corresponding hardware and software used, the associated knowledge for each step, the whole process and needs to ensure 'authenticity' for each step and digital output produced.

We describe below precisely all the full process, describing each main step, and for each step what is being used as input and what is being produced as output as well as the different hardware and software devices that are required.

The description of all the digital files used and/or produced as well as the knowledge and processes is described as RepInfo and is formalized in order to make it make



understandable. All these associated meta-data and descriptions create then the basis to formalize the requirements for OAIS-compliant preservation.

3.3.1 Process details

Adopting a top-down approach for the process pipeline can be represented as shown in the picture below:

The pipeline implicitly defines the data flows: the output of the task_i forms the input of the task_{i+1}. Each stage will be explored in detail in the following pages, but in brief:

- the **Scanning Session** is the sub-task whereby a set of *range-maps* were obtained through several scans of the physical objects. Each scan produces a collection of coordinates for points in the space that represents the object's surface;
- since each *range-map* has its own coordinate system the **Alignment task** aims to combine all the range-maps in one single and common coordinate system;
- the main goal of the **Fusion task** is to join together all these uniform *range-maps* that represent surfaces, in order to produce a 3D model of the solid object.
- finally, a texture can be mapped onto the obtained 3D model. These tasks, entitled **Texture alignment** and **Texture Encoding** are not mandatory.

For each of the above four tasks the following information is provided in detail:

- **task name:** the name of the task;
- **description:** a general description of the specific task in terms of users involved and their main goals;
- **software:** a brief description of the software used;
- **input digital items:** a list of files fed as input into the task and a brief description of their format;
- **output digital items:** the same as input digital items;
- **task information:** general information regarding the specific task with a brief description.

3.3.2 Concrete example for the Testbed

A handcrafted 15 cm-long wooden 'armadillo' was chosen as the test object for scanning and application of the process with the aim of obtaining a digital virtual reconstruction in 3D.

To obtain a cloud of laser points, the armadillo was scanned with a laser scanner MINOLTA VI – 900/910, which is accurate to 0.1 mm. The scanning distance was between 0.6 – 1.2 m while the object lay on a table. Eleven scanning sessions were undertaken. After the completion of the registration process the multiple point sets were aligned, merged and triangulated to build a single, non redundant mesh out of the many, partially overlapping range maps. The output model is available in a Polygon File Format (PLY).

A PLY file consists of a header followed by a list of vertices and then a list of polygons. The header specifies how many vertices and polygons are in the file, and also states



what properties are associated with each vertex, such as (x,y,z) coordinates, normal and colour. The polygon faces are simply lists of indices into the vertex list, and each face begins with a count of the number of elements in each list.

3.3.2.1 *The process flow*

a) **Laser points acquisition**

Task name: Data capture (laser cloud points)

Description: One side or face of the cultural heritage object is scanned.

Hardware: a laser scanner MINOLTA VI – 900/910

- **input:** original cultural heritage object
- **input digital items:** none;
- **output digital items:** a range map in PLY format representing the scanned face of the cultural heritage object;

Task information: general information regarding the specific task with a brief description.

Task information:

- **distance:** 0.8 m from the cultural heritage object;
- **total amount *range-map*:** 11 range maps (laser capture acquisitions, each one corresponding to one side/face of the cultural heritage object) where necessary;
- **plate rotation angle:** none
- **time for acquisition:** 30 minutes.

b) **Geometric registration or alignment of each individual range map so that all can fit together to build the final single range map:**

Task name: Alignment Session

Description: All the different *range-maps* (each with its own geometry and/or coordinate system) have to be transformed in order to obtain *range-maps* with a uniform and common coordinate system. In this task each individual range map is transformed into a new coordinate system (in other words, a matrix transformation).

Software: MeshLab

Input digital items: original PLY files coming from the laser scanner

Output digital items: geometrically corrected / transformed PLY files. RepInfo of the process stored in an ALN file and MA2 file

Task information: The user identifies control points on each pair of the range maps and asks software to transform *range_map1* to match with the geometry of *range_map2*. The same process is applied to each individual range map.

MA2 file describes each individual step.

c) **Combining or fusing each face of the cloud points to obtain a single set of cloud points for the total object (not 11 separated range maps for each face)**



Task name: Fusion Session

Description: All the *range-maps*, with a uniform coordinate system, are combined to produce a 3D model of the object. In addition any redundant data is eliminated, so that each laser point appears only once and represents one point of the surface of the original cultural heritage object. Before this task all the *range-maps* represent a collection of different surfaces that may also overlap rather than form a single solid.

Software: MeshLab

Input digital items: the geometrically corrected set of PLY files and a ALN File

Output digital items: a single PLY file that represents the overall 3D model

Task information:

- **total amount of range-map used:** 11 range-maps
- **precision:** 0.25 mm
- **time needed for the overall fusion task:** 15 min.

d) Data capture for texture (obtaining digital images for each side/face of the cultural heritage object)

Task name: Texture Capture

Description: A series of 8 digital images

Hardware: Canon EOS 350D (digital camera)

Software: none

Input: original cultural heritage object

Input digital items: none

Output digital items: a series of 8 JPG images

Task information:

- **total amount of digital images used:** 8
- **precision:** 72 dpi (resolution)
- **time needed for the overall texture capture task:** 10 min.

e) Merging or aligning texture with 3D model

Task name: Texture Alignment

Description: this task is aimed at geometrically registering the digital images with the 3D PLY. The process requires human intervention where the user identifies control points on the JPG images and their corresponding matching point on the PLY file.

Software: TexAlign, an application developed by the CNR[8]

Input digital items: 8 JPG images

Output digital items: A new PLY file that has the previous 'wire PLY file' plus the TEXTURE. In addition a XML file describing the alignment of all digital images with the 3D model is elaborated.



Task information: The task requires human intervention where the user identifies control points on the JPEG images and their corresponding matching point on the PLY file.

f) Visualizing the 3D model with texture: virtual heritage reconstruction

Task name: Visualization of the 3D model with texture

Description: the 3D models are now textured allow enabling interactive visualization and manipulation for the user.

Software: Virtual Inspector

Input digital items: the single PLY file that represents the overall 3D model

Output digital items: a navigable textured 3D model

Task information:

- **total amount of file:** 1 PLY file
- **time needed for the overall fusion task:** 20 min.



3.4 THE CULTURAL TESTBED SHOWCASE APPLICATION

EWE: an extensible tool for the preservation of VHRP data

Vision

This section describes the main strategy followed for the design of the Cultural Testbed Showcase application. Basically, this application aims to give complete support to people involved in instances of virtual-heritage-reproduction (VHRP) offering the use of CASPAR technologies to preserve their data.

Summarizing, EWE (EWE is an Extensible WEb tool for the digital preservation of VHRP data), addresses the following issues:

- providing an online tool for the definition of vocabularies and schemas,
- offering functionalities for the description (using vocabularies already defined) of projects, activities and digital objects used or produced during a generic VHRP, and
- a transparency layer that allows users to preserve their data in a OAIS-compliant manner using a subset of the CASPAR Key Component.

According to several studies in the UNESCO domain related to testbed activities that clearly define UNESCO's role as a data collector, there is a strong need for a tool that can deal with an enormous degree of heterogeneity coming from several data sources.

UNESCO collects data encoded in different formats and produced with different processes from different data providers. Each data provider uses its own vocabulary, its own methodology and produces files with different software. Note: All this never comes documented to UNESCO, therefore the need to start providing tools for such a documentation. . Just to give a whole overview of the unleashed heterogeneity, consider that the scenario depicted above can be accomplished using several different softwares and the resulting data can be encoded in a group of different of formats.

EWE tries to address these issues providing functionalities that allow users to define their own vocabulary and to describe their projects using the terms contained within. Once these metadata are provided, EWE translates the schema into an RDFS schema extension of the CIDOC-CRM Ontology and then converts the metadata into instances of such schema. This process aims to produce DescInfo and RepInfo to be packed in a OAIS AIP.

The following document provides a minimum overview of the EWE internal data model, its architecture, a basic use case and an implementation plan.

Data model

The main aim of the *Conceptual Model for the XML Project Description Language Specification* [10] document is to describe the conceptual model behind the XML Project Description Language. This latter is an XML language for the internal



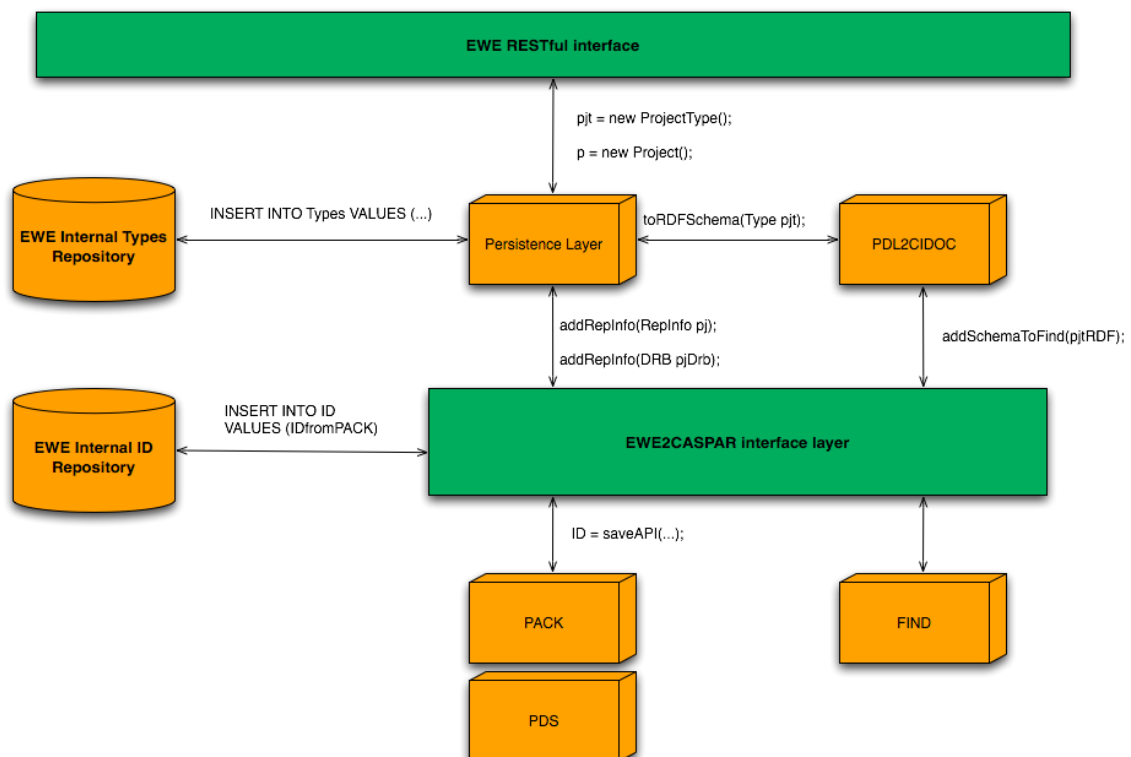
representation of projects and activities to be handled with EWE. Following a bottom-up approach all the main entities are defined and an informative example is also shown.

It has been decided to not include here such specifications just for the sake of simplicity. Please refer to the document and presentation provided at the Cultural Testbed Authenticity Meeting [ref] for more details.

The key concept here is that projects can be described in terms of their activities and their digital objects, and such descriptions (encoded in a XML) translated into an RDFSchema that is an extension of the CIDOC-CRM ontology.

Architecture

The following picture shows the EWE high-level architecture. Each component is therefore described.



The **EWE RESTful interface provides** a set of RESTful APIs allowing users to define project types. Every project type is described by a set of activities each dealing with a type of digital object (i.e. files). Since the final aim of the application is to produce data that can be translated into several RepInfos and DescInfos, the concept of Digital Object Type is the main point here.

In fact, since a Digital Object Type is an XML document describing the format of a specific file with others data related to the production and to context of it, its translation



in RDF make possible to store a DescInfo within an AIP in easy way. The feasibility of achieving this using XSPARQL[11] is currently under discussion.

Moreover, a Digital Object Type, embeds all the structural information (more specifically Structural RepInfos) needed for the long-preservation term in a low coupled way: in fact, a separated and specific DRB File is attached to every Digital Object Type. This component also acts as a transparency layer that hides all the OAIS technologies that underly the application. Furthermore, the choice of exposing the application functionalities as a REST service allows us to build several different Web applications on top achieving a high degree of extensibility.

EWE Core

This component has the application business logic responsibility. Basically, it catch user input, stores and retrieves types and instances in a relational database, converts them into XML to be provided as output of the REST APIs, sends them to the component for the RDFSchemata/RDF translation and, finally, is responsible for sending the overall data for CASPAR ingestion/retrieval.

PDL2CIDOC

This is the component responsible for converting instances of the two EWE internal languages into a formalism to be sent to the CASPAR Find component as DescInfo. More specifically:

- An instance *A* of the EWE XML Project Description Language was converted into one RDFSchemata A_{RDF} that is an extension of the CIDOC-CRM Ontology while,
- An instance *a* of the EWE XML Project Instance Description Language (that represents a concrete project compliant to the description *A*) was converted into an instance of the A_{RDF} schema.

EWE Internal Types/Instances Repository

Mainly, a relational database where types and instances are stored.

EWE2CASPAR Interface Layer

This module acts as a proxy between the concrete CASPAR Key Components (more specifically PACK and FIND) and the EWE internal data representation. It also provides storage in a relational database (**EWE Internal ID repository**) for the identifiers obtained from the PACK component in order to fulfil retrieval requirements.

PACK, **PDF** and **FIND** will obviously not be described.

Implementation Plan



Objectives(s)	Analysis and study of the UNESCO scenario and the overall cultural domain		
Outcome(s)	Several datasets with an analysis of suitable preservation requirements, in terms of RepInfo available/needed.		
Tasks			
Start date	April 2008	End Date	September 2008
Status	CLOSED		
Concrete Results	Villa Livia dataset analysis and description. See <URL> for details.		

Objectives(s)	Identification of a suitable scenario to be used for the testbed validation		
Outcome(s)	A fully described and documented scenario in terms of main processes, actors, activities and their preservation needs.		
Tasks			
Start date	August 2008	End Date	October 2008
Status	CLOSED		
Concrete Results	The 3D Laser scanning process (as it is currently implemented as the CNR ISTI VLab) is completely described and its preservation needs assessed. An example of its dataset was obtained. See <URL> for details.		

Objectives(s)	EWE Prototype development		
Outcome(s)	A proof-of-concept prototype of the EWE main functionalities. It shows schemas definition and instantiation (and their storage/retrieval) through an AJAX web application.		
Tasks			
Start date	October 2008	End Date	November 2008
Status	CLOSED		
Concrete Results	EWE Prototype		



Objectives(s)	EWE Component development: EWE Core		
Outcome(s)	Development of the EWE Core component		
Tasks			
Start date	November 2008	End Date	December 2008
Status	IN PROGRESS		
Concrete Results			

Objectives(s)	EWE Component development: EWE RESTful API		
Outcome(s)	EWE RESTful API		
Tasks			
Start date	December 2008	End Date	January 2009
Status	IN PROGRESS		
Concrete Results			

Objectives(s)	EWE Component development: PDL2CIDOC		
Outcome(s)	EWE PDL2CIDOC Component		
Tasks			
Start date	December 2008	End Date	January 2009
Status	TO BE STARTED		
Concrete Results			

Objectives(s)	EWE Components integration with CASPAR PACK and FIND		
Outcome(s)	an integrated version of EWE		
Tasks			
Start date	February 2009	End Date	March 2009
Status	TO BE STARTED		
Concrete Results			



Objectives(s)	EWE Functional testing and deploying		
Outcome(s)	a deployed and publicly available version of EWE		
Tasks			
Start date	March 2009	End Date	March 2009
Status	TO BE STARTED		
Concrete Results			

3.5 VILLA LIVIA DATASET OVERVIEW

This page contains all the information about [Villa Livia](#) Cultural Testbed Data, their RepInfo and DescInfo. Collected files have to be intended as an example of digital heritage data obtained as laser range scans, GPS data or traditional archaeological documentation.

The Villa Livia dataset is a collection of files used within the "[virtual museum of the ancient Via Flaminia](#)" project: a 3D reconstruction of several archaeological sites along the ancient Via Flaminia, the largest of them being *Villa Livia*:



A rough estimate of the total dataset size is 500 GB. File types in this set include:

- 3D point clouds (imp, dxf, dwg)
- Elevation grids (agr, bt)
- 3D meshes (mdl, vrml, v3d)
- Textured 3D models (max, pmr, ive, osg)
- Satellite data (ers, ecw)
- GPS data, maps (txt, apm, shp)
- Digital images (targa, jpeg, tiff, png, psd, bmp, gif, dds)

Currently (as of end of Y2), two file types have been used for testing:



- an **elevation grid** of the site (agr / grd)
- a **map** of the site contours (shp)

3.6 ESRI ASCII GRID FILE: *DEM_LOD3_LIVIA.GRD*

This is an elevation grid (height map) of the area where *Villa Liva* is located. It is an ASCII file in the [ESRI GRID file format](#):



3.6.1 Structural and Semantic RepInfo for ESRI GRID file format.

DataObject [dem_LOD3_livia.grd](#) and related RepInfo relationship:



RepInfo relationships" width="638" height="233" />

where

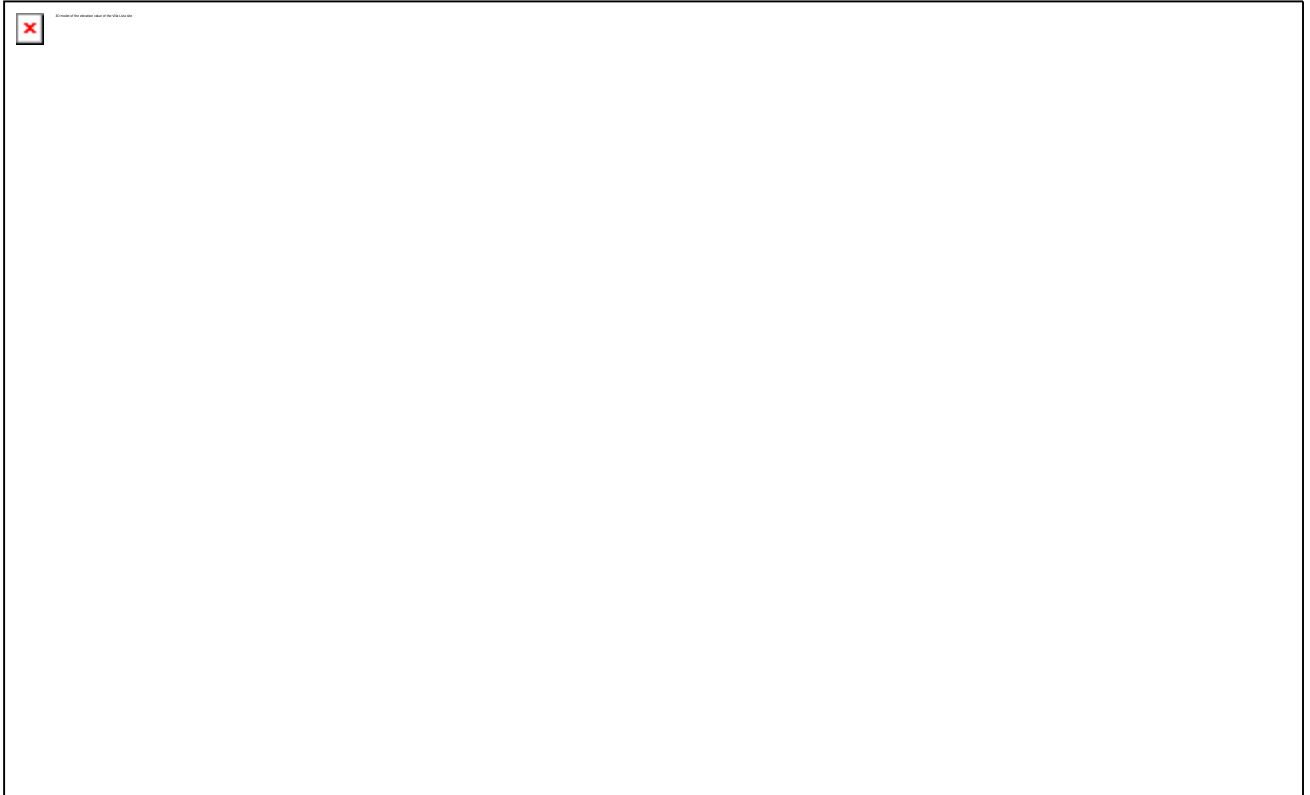
- [esri_ascii_grid.xsd](#) is the XML schema describing the ESRI ASCII GRID File to be used with the [Data Request Broker](#) tool. It provides information about the structure of the DataObject.
- [sdf-20020222.xsd](#) is the XML schema of the **Structured Data File** implementation. It defines XML elements to be used in the `esri_ascii_grdi.xsd` schema.
- <http://orlando.drc.com/SemanticWeb/DAML/Ontology/GPS/Coordinate/ver/0.3.6/GPS-ont#> is the XML namespace of the **DAML ontology for GPS coordinate values**, adding meaning to *xllcorner* and *yllcorner*.
- [ESRI GRID file format specification](#)
- Data Request Broker: Structured Data File implementation notes. SDF breaks down any binary file into a tree of nodes thanks to an external description. The internal description is an XML Schema with a few additional markups providing the physical description of the binary file. [drbdemo for ESRI ASCII Grid.zip](#) a DRB demo example with Shape file data can be found.
- [esria_ded.xml](#) is an instance of the Data Entity Dictionary Specification Language. It allows to add some simple data semantics.



- [dedsl.xsd](#) is the XML schema for the Data Entity Dictionary Specification Language. See [DEDSL Schema page](#) for more information.

3.6.2 Preservation Description Information for an ESRI GRID file AIP

The picture below represents the complete AIP for ESRI GRID files:



where:

- **Provenance:** [villa livia dem LOD3 livia.rdf](#) is the RSLP collection description created with the online tool available at [Research Support Libraries Programme](#). This file describes a collection, its location and associated owner(s), collector(s) and administrator(s).
- **reference:** TO BE DONE
- **Context:** TO BE DONE
- **Fixity:** TO BE DONE

3.6.3 Descriptive Information

TO BE DONE

3.6.4 Complete AIP for ESRI GRID file

[UNESCO Villa Livia 20080501_AIP_V1_1.zip](#) First Draft Elevation Grid data AIP built using PACK Component.



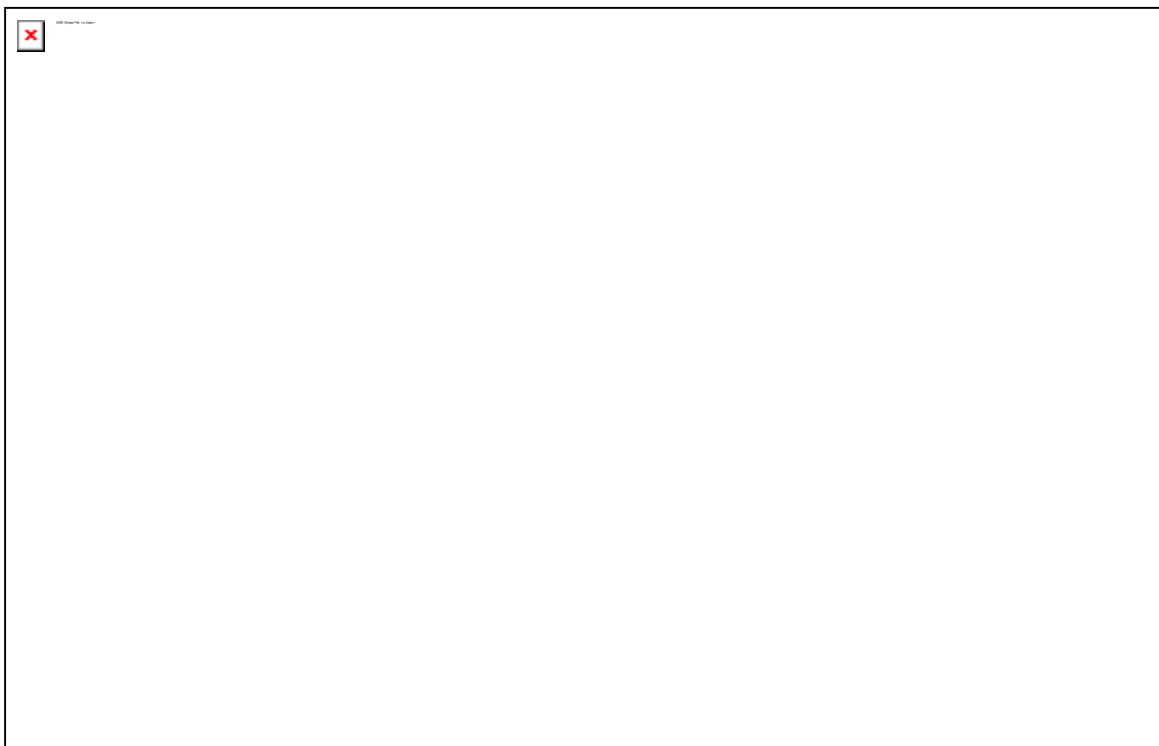
3.7 ESRI SHAPE FILE: *VINCOLI_LIVIA.GRD*

It is a vector file of site contours. It is a binary file in the [ESRI Shape file format](#). A possible visualisation:



3.7.1 Structural and Semantic RepInfo for ESRI Shape file format.

DataObject [vincoli livia.shp](#) and related RepInfo relationship are showed in the figure below:



RepInfo relationships" width="638" height="233" />

where:



- [esri_shapefile.xsd](#) is the XML schema describing the ESRI ASCII GRID File to be used with the [Data Request Broker](#) tool. It provides information about the structure of the DataObject.
- [sdf-20020222.xsd](#) is the XML schema of the **Structured Data File** implementation. It defines XML elements to be used in the esri_shapefile.xsd schema.
- [ESRI Shape file format](#)
- Data Request Broker: Structured Data File implementation notes. SDF breaks down any binary file into a tree of nodes thanks to an external description. The internal description is an XML Schema with a few additional markups providing the physical description of the binary file.
[drbdemo_for_ESRI_SHAPEFILE_advanced.zip](#) an DRB demo example with Shape file data can be found.

3.7.2 Preservation Description Information for an ESRI Shape file AIP

NOTE:AIP creation for Shape file is still in progress

The picture below represents the complete AIP for ESRI Shape files.

TO BE DONE

Where:

- **Provenance:** TO BE DONE
- **reference:** TO BE DONE
- **Context:** TO BE DONE
- **Fixity:** TO BE DONE

3.7.3 Descriptive Information

TO BE DONE

3.7.4 Complete AIP for ESRI Shape file

TO BE DONE

3.8 RELATED DOCUMENTATION

- [UNESCOJune19CasparreviewJune2008v14.ppt](#) Cultural Testbed presentation on the June 19 2008 EU review
- [GRID_AIP.pdf](#) ESRI ASCII AIP Overview

3.9 OTHER MISC DATA WITH A BRIEF DESCRIPTION

- [ESRI_wikipedia.zip](#): archived wikipedia page ESRI grid format



- [ESRI_wiki_shapefile.zip](#): ESRI shapfile format archived wiki page
- [html40.txt](#): HTML 4.0 specification in plain text
- [ISO-IEC-14772-VRML97.zip](#): ISO standards for VRML
- [msn-dds_format.txt](#): TEXT file containing link to MSN format support for DDRS
- [shapefile.pdf](#): White paper on shapefiles
- [VRML97Am1.zip](#): ISO extension to VRML standard adds geospatial NURBS

3.9.1 UNESCO Glossary

VHRP - Acronym for Virtual Heritage Reconstruction Processes. The class of all the processes aimed at the digital reproduction of physical and existent cultural heritage.

CASPAR-based application - an application that uses of at least one of the components developed within the CASPAR Project in order to achieve some digital preservation needs.

Range-map - A Range Map is a two-dimensional image, where each pixel is the floating point distance from the image plane to the object in the scene. This is especially useful for generating synthetic data sets for use in Computer Vision research, e.g. depth from stereo and shape from shading.

3.9.2 UNESCO References

[1] <http://www.casparpreserves.eu>

[2] UNESCO, *Convention concerning the protection of the world cultural and natural heritage*, 17th Session General Conference, 16 November 1972, Paris

[3] <http://www.vhlab.itabc.cnr.it/flaminia/>

[4] Maurizio Forte presentation at 'Using New Technologies to Explore Cultural Heritage Conference', 5 October, 2007

[5] Patrick Min, 'CASPAR Cultural Testbed Scenarios', available on the CASPAR wiki

[6] <http://wiki.casparpreserves.eu/bin/view/Main/CulturalTestbedArchitecture>

[7] P. Min, D. Palmisano, M. Hernandez, 'Cultural Testbed Presentation', CASPAR Project

24 Review, 24th-25th June 2008, Brussels

[8] <http://vcg.isti.cnr.it/> Visual Computing Lab informal metadata for the scanning archiving

[9] PLY File Format <http://local.wasp.uwa.edu.au/~pbourke/dataformats/ply/>

[10] http://wiki.casparpreserves.eu/pub/Main/EWE/conceptual_model.pdf



[11] <http://axel.deri.ie/xsparq>



4 CONTEMPORARY PERFORMING ARTS TESTBEDS

This section contains details of the testbeds in each of the partner organisations with contemporary arts data. The question of authenticity is of particular interest since we are not concerned usually here with the recording of audio or video but rather the re-performability of a composition.

4.1 IRCAM TESTBED

Authors: Jerome Bartholemy

4.1.1 Test cases

Ircam testbed will be based on the set of test cases currently available in the Mustica server, that is to say, musical works with digital components. Currently, there are 72 musical works available in the current repository.

In a first phase, the tests will be applied to a selection of these.

In a second phase, the whole set of musical works will be migrated to MustiCASPAR, the version of MUSTICA which has CASPAR features incorporated.

It shall be noticed that this operation can be made automatically, but in this case all the knowledge associated to objects should be insufficient. Thus, more RepInfo and PDI should be generated for the objects.

4.1.2 Test scenarios

The scenarios will be based on the scenarios already provided in the D4101 CASPAR deliverable.

The scenario developed here is related to:

**Scenario 3 (D4101 – chapter 4.1.4.1):
changes in the operating system software, changes in the hardware, or changes in the real-time processing engine (e.g. version upgrade).**

The changes can affect the hardware or the operating system where the real-time software (e.g. MAX/MSP) is running. The same scenario is applicable with changes that can affect the real-time software (e.g. MAX/MSP).

The questions are:

1. does the changes affect the behaviour of the patch?
2. does the CASPAR process help identify which behaviour are affected by the changes?
3. does the CASPAR process help identify if these behaviours affect the logic of the work (the "patch" itself)?

The CASPAR process can help solving these issues by investigating the development of a model for the description for the logic of the work (the patch), a model of description for the real time process, and develop a mechanism for identifying affected behaviours. A mechanism for helping to identify changes in perceived results, and preserving authenticity is also needed.

**Scenario 4 (D4101 – chapter 4.1.4.1):
changes in the availability of the real-time processing engine.**



The changes can affect the the real-time software which is no longer available.

The questions are:

- does the CASPAR process help identify an equivalent software where the patch can be implemented?
- does the CASPAR process help identify how the new implementation affects the logic of the work?

The same considerations apply as for Scenario 3.

SCENARIO SUMMARY

- We (as archivists) ingest of a new WORK into MustiCASPAR,
- We (as registered experts) receive an asynchronous notification of loss of availability for a COMPONENT,
- We (as registered experts) search for equivalent COMPONENTS,
- A new version of the COMPONENT becomes available,
- We (as archivists) ingest of the new version of the COMPONENT.

USED PACKAGES

- Representation Information Toolbox
- Registry
- Authenticity manager
- Preservation Orchestration Manager
- Packaging
- Knowledge Manager
- Preservation Data Store



SUB-SCENARIO #1

Ingest of a new WORK

STEPS
Collect DATA for the WORK
For each COMPONENT of the WORK {
Search the REGISTRY for existing CPID using features
Chose best compliant existing RepInfo CPID
If no one CPID is found {
Register / Log into RepInfo toolbox portal
Use Analysis tools to produce RepInfo for the COMPONENT
}
Else {
Retain and use CPID
}
Analyze DATA object and produce AP and PDI objects
Construct AIP
Store AIP
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for meta-data added
Store DATA into DATA STORE
}
Analyze WORK and produce AP and PDI objects
Construct AIP of the WORK
Store AIP of the WORK
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for meta-data added

SUB-SCENARIO #2

Notification of loss of availability for a COMPONENT

STEPS
Register / log into PRESERVATION ORCHESTRATION MANAGER (POM)
Search the REGISTRY for the old existing CPID (CPID _{old})
If CPID is not found {



Register / Log into RepInfo toolbox portal
Use Analysis tools to produce RepInfo for old existing COMPONENT
Store RepInfo into REGISTRY
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for change of RepInfo
}
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for loss of availability of the COMPONENT



SUB-SCENARIO #3

Search for equivalent COMPONENTs

Strategy 1 – Search on the basis of provenance (all porting of existing component)

STEPS
Search the KM for existing CPID using provenance
If no one CPID is found {
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for loss of availability of the COMPONENT
}
Else {
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for availability of new version for the COMPONENT
}

Strategy 2 – Equivalent component on the basis of ontological RepInfo

STEPS
To be defined later if needed

External porting of obsolete COMPONENT

SUB-SCENARIO #4

Ingest of a new version of a COMPONENT

STEPS
Receive the ALERT from the PRESERVATION ORCHESTRATION MANAGER (POM) containing the CPID of existing COMPONENT (CPID _{old}), CPID of a new version of a COMPONENT (CPID _{new}) and the AUTHENTICITY PROTOCOL (AP _{new}).
Search the REGISTRY for the old existing CPID (CPID _{old})
If CPID is found {
Retrieve the associated AUTHENTICITY PROTOCOL (AP _{old})
If AP exists {
Verify existence, consistence and correct application of AUTHENTICITY PROTOCOL (AP _{old})
If AP is “equivalent” {
Search all the WORKS, belonging to the managed DATA STOREs, using the COMPONENT (CPID _{old}).



For each WORK using COMPONENT (CPID _{old}) {
Retrieve existing ARCHIVAL INFORMATION PACKAGE (AIP)
Retrieve DATA of new COMPONENT (CPID _{new})
Ingest a new WORK after substituting the old COMPONENT (CPID _{old}) with new COMPONENT (CPID _{new})
}
Notify the PRESERVATION ORCHESTRATION MANAGER (POM) for availability of new version for the COMPONENT
}
Else {
Notify PRESERVATION ORCHESTRATION MANAGER (POM)
}
}
Else {
Notify PRESERVATION ORCHESTRATION MANAGER (POM)
}
}
Else {
Notify PRESERVATION ORCHESTRATION MANAGER (POM)
}
}

Test phases

Ircam testbed implementation is planned in four different phases :

- The first phase is for definition and setup of the infrastructure.
- The second phase is for training of users and production of adequate documentation.
- The third phase is for definition of test scenarios and criteria for validation, as well as providing support for validation (questionnaires...)
- The fourth phase is for validation itself, collection and analysis of user's feedback, and production of relevant recommendations and outcomes.

4.1.3 Definition and setup of testbed infrastructure

Objective(s)	To develop an adequate infrastructure for CASPAR test and validation
--------------	--



Outcome(s)	<ul style="list-style-type: none"> • A server and a front end able to store (ingest) and retrieve (access) musical works and providing access to the CASPAR services 		
Tasks	<ul style="list-style-type: none"> • Choice of the adequate infrastructure • Integrate and test CASPAR Key Components • Develop specific User Interface 		
Start Date	M18	End Date	November 08

4.1.4 - Training, production of documentation

Objective(s)	To train future users for using the infrastructure and implementing scenarios.		
Outcome(s)	<ul style="list-style-type: none"> • A set of tools for user's support (documentation) 		
Tasks	<ul style="list-style-type: none"> • Selection of testing staff • Production of user's support : Video capture, Written documentation, Online documentation • Organization of specific training events 		
Start Date	December 08	End Date	February 09

4.1.5 - Definition of scenarios and validation support

Objective(s)	To develop test scenarios on the basis of the already described scenarios (D4101), and define support tools for validation.		
Outcome(s)	<ul style="list-style-type: none"> • Detailed scenarios ready for implementation and testing • Support for validation (questionnaires, methodology, validation criteria) 		
Tasks	<ul style="list-style-type: none"> • Selection of test cases • Refinement of scenarios (definition of use cases) • Development of support for validation : questionnaires, validation criteria 		
Start Date	December 08	End Date	February 09

4.1.6 – Tests and validation

Objective(s)	To implement test scenarios and validate CASPAR services		
Outcome(s)	<ul style="list-style-type: none"> • An analysis of the uses of CASPAR components and services • A set of recommendations for implementation of CASPAR services 		



Tasks	Collection of users feedback Analysis of users feedback Elaboration of recommendations		
Start Date	March 09	End Date	June 09



4.2 UNIV LEEDS TESTBED

Authors: Kia Ng, Eleni Mikroyannidi

4.2.1 Test cases

4.2.1.1 Demo data and Representation Information

<http://wiki.casparpreserves.eu/bin/view/Main/UNIVLEEDSTestbedRepInfo>

4.2.1.2 DC Profiles

Composer/Performer

This profile covers the simple user who is not an expert with the technologies used for the 3d motion capture or with the technologies related to the archival system. This user can use the system with the assistance of a Multimedia expert or Developer.

Multimedia User Profile

This profile describes the multimedia expert who is mainly working on the capture of the performance by the vicon system. This user is familiar with the format and content of the Demo data that are captured.

Developer

A work is ingested in the repository by the Developer/Operator expert. The developer/Operator is familiar with the OAIS model and the technology of the archival system.

4.2.2 Scenarios

It should be noted that according to the different versions of the ICSRiM Archival System, the procedure of implementing the scenarios differs. The general description of the scenarios follows.

4.2.2.1 Scenario 1: Creation of a new Information Package

Description

A Composer/Performer has set an appointment with either an Operator or a multimedia User in order to record his/her performance with the vicon 3d motion capture system. This performance should be described and kept for future use in the ICSRiM repository. After the capture, the Operator will describe and ingest the work in the archival system.

Steps

- The Performer with the help of the multimedia user records his/her performance
- The generated performance data are kept on a temporary place
- The Operator describes the performance using the RepInfo tool (DC member tool). A new RepInfo object is generated and exported in rdf format.
- The Operator imports in the archival system the RepInfo object.
- The archival system detects the modules (PDF, MOV etc) and asks from the operator to complete the package by adding the corresponding performance data.



- The Operator adds the performance data and submits the whole package into the repository
- The ingestion of the information package has completed

4.2.2.2 Scenario 2: Retrieve an Information Package

Description

The user wants to access past performances that have been ingested in the archival system.

Steps

- The user logs into the archival system
- He/she searches for the performance
- The archival system executes sql queries into the RepInfo objects and returns the related results into a table
- The user will be able to download the various files he/she is interested in. In addition, he/she will be able to see the description and details of an Information Package with the RepInfo tool.

4.2.2.3 Scenario 3: Updating an Information Package

Description

The developers of the MAX/MSP patch (AMIR s/w), which is responsible for the recording and recreation of the performance, have come up with a new version. This new version is related with the following files:

- SDIF: Data generated by Vicon iQ2.5. It includes both sound (AIFF) and TRC data into a single file
- CFG: A new configuration file for the new AMIR version. It replaces the old CFG file.
- MXT: The new format of the MAX/MSP patch. It replaces the old MXF format

The new AMIR version supports more features during the capture. A task is to update an old Information package in order to make it compatible with the new amir version. The following steps are followed:

- Transformation of an old package in order to be processed by the new s/w
- Update the archival system with the new information

Different strategies can be followed in order to achieve the previous task. However, the most convenient solution will be the preservation of all the data and their various versions. Thus, the Information Package is enriched with the files that support the new version of the Patch. The archival system has to be updated with the new information. In order to achieve this, the following strategies are followed:

4.2.2.3.1 Strategy 1 (update the old package):



- The user follows the procedure of scenario 2 in order to retrieve the Information Package he/she wants to update.
- The user retrieves the RepInfo object that describes the Package he/she wants to update
- The user updates the existing RepInfo object with the RepInfo tool (DC member tool). He/she updates the RepInfo with the new formats and file description.
- The updated RepInfo object (exported in RDF) is checked by the archival system, which discovers the files that are missing (sdif, mxt, cfg).
- The system gives a notification about the new formats that are not defined in the modules (sdif, mxt)
- The user updates the GapManager? with the new modules (sdif, mxt)
- The archival system asks from the user to complete the package by adding the files that are described in the rdf file (sdif, mxt, cfg).
- The Information Package has been updated

4.2.2.3.2 Strategy 2 (create a new package and relate it through comments with the old one):

- The user follows the procedure of scenario 2 in order to retrieve information about the Information Package he/she wants to update.
- The user follows the procedure of scenario 1 and creates a new Information package in order to create a new Information package containing only the updates of the old Information Package
- During the creation of the new RepInfo object, he/she adds a comment relating it with the old Information Package
- A new Information Package has been created containing the updated files

4.2.3 Testbed Phases

The Univleeds artistic testbed implementation plan contains 5 main phases, according to the added functionality:

- Phase 1 includes the definition and setup of the testbed infrastructure
- Phase 2 includes the definition of the preservation scenarios
- Phase 3 includes the integration of the key components
- Phase 4 includes the training and production of documentation
- Phase 5 includes the implementation of the preservation scenarios and validation of the CASPAR services

According to the overall implementation plan, the following Gantt diagram is provided:



4.2.3.1 Phase 1: Definition and setup of testbed infrastructure

Objective(s)	<ul style="list-style-type: none"> The setup and development of the Univleeds testbed infrastructure 		
Outcome(s)	<ul style="list-style-type: none"> A simple web archival system that will consist of a web interface and backend for implementing the preservation scenarios 		
Tasks	<ul style="list-style-type: none"> Choice of infrastructure Setup server - MySQL connectivity User authentication/registration Development of the web interface using JSP in Eclipse IDE environment ReplInfo construction for different performances 		
Start Date	M22	End Date	M32

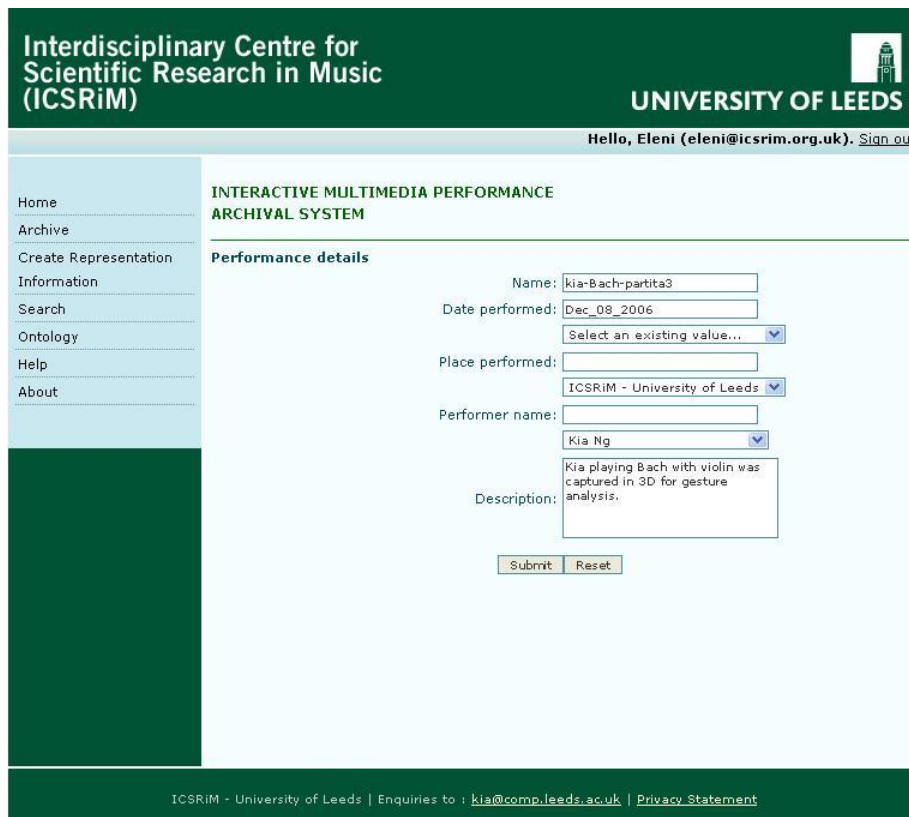


Figure 1: Creation of the Representation Information with the use of the archival system

4.2.3.2 Phase 2: Scenario Definition

Objective(s)	<ul style="list-style-type: none"> Define test scenarios covering the requirements for preservation 		
Outcome(s)	<ul style="list-style-type: none"> Scenarios based on the preservation requirements Detailed and concise use cases for testing and implementation 		
Tasks	<ul style="list-style-type: none"> Definition of preservation requirements for our testbed Definition of scenarios Definition of detailed use cases 		
Start Date	M25	End Date	M35



4.2.3.3 Phase 3: Key Components Integration

Objective(s)	The integration of CASPAR components with the ICSRiM archival system. This will be achieved through web services, as this is the interface that the CASPAR components are expected to implement.		
Outcome(s)	<ul style="list-style-type: none"> • An archival system that can be used for implementing the preservation scenarios. • Feedback on the key component integration 		
Tasks	<ul style="list-style-type: none"> • Survey on the available components • Selection and tests on key components. According to the specified scenarios and the survey, the most suitable components will be selected for integration with the archival system. • Web services integration 		
Start Date	M22	End Date	M39

4.2.3.4 Phase 4: Training

Objective(s)	The selection and production of testing material and support documentation		
Outcome(s)	<ul style="list-style-type: none"> • Training material for the support the archival system 		
Tasks	<ul style="list-style-type: none"> • Selection of testing material • Production of support such as written documentation, video, screen captures etc. • Organization and participation in training events 		
Start Date	M33	End Date	M41



4.2.3.5 Phase 5: Tests

Objective(s)	Implementation of different use cases with the use of the archival system and validation of the CASPAR services.		
Outcome(s)	<ul style="list-style-type: none"> • Tests analysis on the CASPAR services • Feedback on the implementation process and the CASPAR services 		
Tasks	<ul style="list-style-type: none"> • Tests analysis and scenario implementation • Revision of the support documentation 		
Start Date	M35	End Date	M41



4.3 CIANT TESTBED

Authors: Michal Masa

Test cases

The final version of CIANT testbed should contain a set of works from the field of **contemporary art and cultural heritage**:

- New media performances
- Video art
- Digitized 3D models

Specific instances of these categories will be:

- Golem (New media performance)
- V.I.R.U.S. (New media performance)
- GAMA (Video art)
- Langweil's model of Prague (Digitized 3D models)

Test phases

- Phase 1 – Definition and setup of CASPAR infrastructure
- Phase 2 – Definition and setup of DC tools
- Phase 3 – Training and production of documentation
- Phase 4 – Definition of scenarios
- Phase 5 – Test and validation through user feedback and outcomes

Phase 1 Objective(s)	Setup of CIANT testbed structure		
Outcome(s)	Working installation of Tiger release on CIANT facilities		
Tasks	Setup of the components, integration tests, communication with remote installations of Knowledge Manager and Registry		
Start Date	M18	End Date	December 08



4.4 INA-GRM TESTBED

Authors: Michael Gatt, Yann Geslin

Test cases

The INA-GRM testbed will be based on a number of different acousmatic pieces. Each acousmatic piece will contain similar data objects, but the creation of these objects will have been achieved in a variety of different way. A selection of these acousmatic pieces will be used to apply tests, after which more pieces will be gathered to ingest into MustiCASPAR. As INA will be using MustiCASPAR much of the phases will be similar to that of IRCAM's.

Test phases

- Phase 1 – Definition and setup of the infrastructure.
- Phase 2 – Training and production of documentation.
- Phase 3 – Definition of scenarios.
- Phase 4 – Test and validation through user feedback and outcomes.

Phase 1

Objective(s)	Setup of INA-GRM testbed structure		
Outcome(s)	<ul style="list-style-type: none"> • Designing of DC member tool in conjunction with CNRS • Working alongside IRCAM to insure compatibility with MustiCASPAR 		
Tasks	<ul style="list-style-type: none"> • Choice of an adequate infrastructure • Integrate and test with MustiCASPAR (which adheres to Key Components) • Design user interface for DC Member tool alongside CNRS 		
Start Date	M18	End Date	November 08

Phase 2

Objective(s)	To train and inform users		
Outcome(s)	<ul style="list-style-type: none"> • Documentation on DC member tool use and integration within MustiCASPAR 		
Tasks	<ul style="list-style-type: none"> • Selection of testing staff • Produce documentation for users and INA specific documentation 		
Start Date	December 08	End Date	February 09

Phase 3

Objective(s)	Create test scenarios		
Outcome(s)	<ul style="list-style-type: none"> • Detailed scenarios for implementation and testing 		
Tasks	<ul style="list-style-type: none"> • Further selection of acousmatic works • Refinement of works with scenarios 		
Start Date	December 08	End Date	February 09



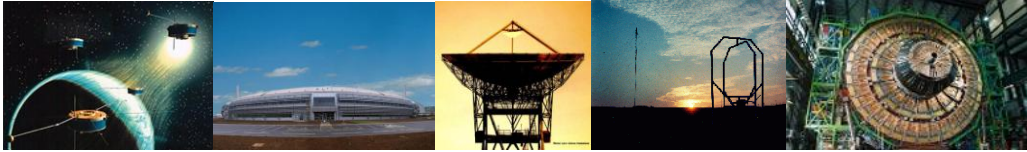
Phase 4

Objective(s)	Use scenarios to validate CASPAR services		
Outcome(s)	<ul style="list-style-type: none"> • Recommendations for implementation of CASPAR services • Analysis of CASPAR components and services 		
Tasks	<ul style="list-style-type: none"> • Collect user feedback • Analysis user feedback • Make recommendations 		
Start Date	March 09	End Date	June 09



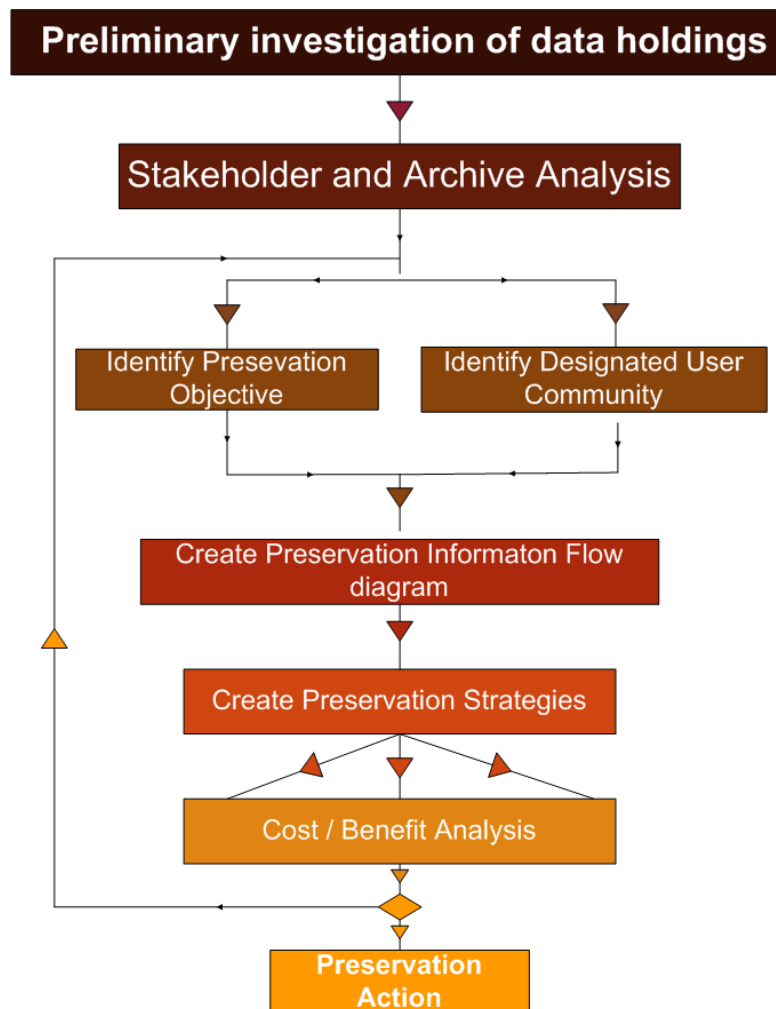
5 SCIENCE TESTBEDS

5.1 STFC TEST BEDS



For the STFC testbeds a methodology was developed in response to the challenge of digital preservation. This challenge lies in the need to preserve not only the dataset itself but also the ability it has to deliver knowledge to a future user community. The preservation objective is defined by the knowledge that a data set is capable of imparting to any future designated user community and has a profound impact on the required preservation actions an archive must carry out.

We sought to incorporate a number of analysis techniques tools and methods into an overall process capable of producing an actionable preservation plan for scientific data archives. The workflow below illustrates the stages of this analysis methodology (Full details of the methodology are available in a separate document)





5.1.1 OAIS Preservation information flow diagrams

For each of the test bed scenarios an OAIS preservation information flow diagram was created. It is graphical representation and analysis tool which is a hybrid of information flow diagram and the OAIS reference model.

Elements of OAIS Preservation information flow diagrams

Standard OAIS reference model components of an AIP. These are the standard components of an AIP. All information entities must be mapped to at least one of the following component within an AIP.

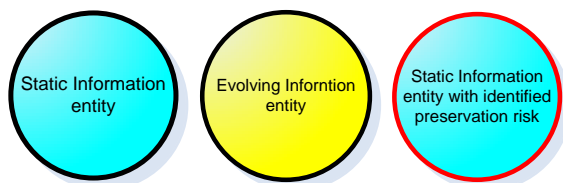
- Content
- Representation Information
 - Structure
 - Semantic
 - Other
- Preservation Description Information
 - Reference
 - Fixity
 - Context
 - Provenance

5.1.2 Information Objects

An information object is a physical unit of information suitable for deposit within an AIP as it currently exists. An information object must have the following attributes

- Name
- Description of information contained by entity which is vital for the preservation objective e.g. a piece of software contains structural information and algorithms for the processing of data within it's code
- Description of format i.e. website, PDF, database or software
- Assessment of preservation risks and dependencies

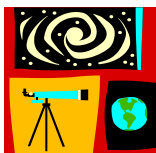
Notation used



Stakeholder entities

A stakeholder entity is the named custodian of the required Information entity

Notation used



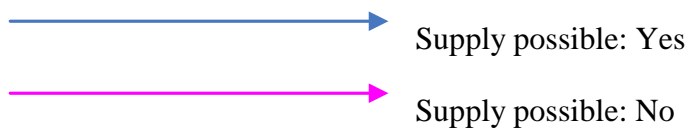


5.1.3 Supply Relationship

The supply mechanism should simply be an indicator of any impediment to the current supply of an information entity such as an embargo or assertion of copyright. The attributes of a the supply relationship are

- Supply possible (Yes/No)
- Description of supply impediment

Notation used



5.1.4 Supply Process

The supply process is any process carried out on information supplied by the stakeholder in order to produce the information object. Its attributes are

- Name
- Description of process e.g. dump of a database table into a csv file, archiving of public website or reformatting of data files

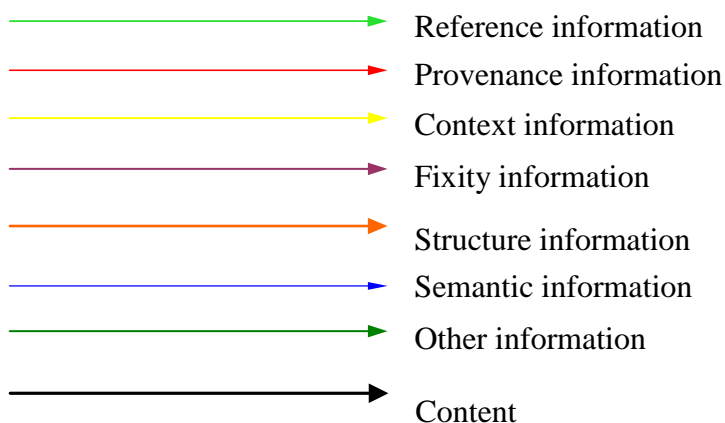
Notation used



5.1.5 Packaging relationship

The only required attribute of the packaging relationship is that it links an Information entity to at least one standard OASIS reference model component of an AIP. However many implementations of packaging such as XFDU require additional information.

Notation used





5.1.5.1 Information object dependency relationships

The information object dependency relationship connects two information objects. If preservation action is carried out on one object the impact on another object with a dependency. For example if a piece of software is identified to be at preservation risk and deconstructed to a structural format and analysis algorithm descriptions, the software user manual will be flagged up by the dependency relationship and may be removed on the basis that this information is now irrelevant.

Notation used

5.1.6 Preservation strategies

The Information flow diagram should now identify where preservation strategies need to developed in following areas.

5.1.6.1 In response to a supply impediment

Where there is an impediment to the supply. A strategy must be developed in order to either overcome the impediment immediately for example purchasing a special licence for software or an institution could develop a simplified open source version of the software which contains the key functionality. The alternative is to develop a mechanism that effectively references the external information object in tandem with a mechanism for monitoring the situation (preservation orchestration).

5.1.6.2 In response to an identified information preservation risk

Information objects must be inspected on a case by case for their individual preservation risk based on dependencies they have which will be affected by the passage of time. Different strategies which effectively obviate these risks must then be developed.

5.1.6.3 As a secondary response to a preservation strategy

Where a dependency between information objects have been identified secondary preservation strategies may need to be developed for a related objects.

Multiple strategies can be developed for each instance in these areas. This results in a number of preservation plans being formed.

A preservation plan consists of a unique

- Set of information objects
- Set of supply relationships
- Set of preservation strategies

Which allow you to carry out a series of clear actions in order to create an AIP. This allows you to take a number of plans to the cost/benefit stage.

Cost/Benefit Analysis

Plan options can then be assessed according to

- Costs to archive directly as well as the resources knowledge and time of archive staff
- Benefits to future users which ease and facilitate re-use of data



- Risks – what are the risks inherent the preservation strategies and are they acceptable to the archive.

5.1.7 The Implementation Plans

The Implementation Plan we present here is a brief summary of our intended actions containing the information flow diagrams created in the analysis phases and required actions based on selected preservation strategies for 4 scenarios on two different scientific data archives. Full discussions of the information objects and available strategies will be written up as separate case studies

5.1.7.1 Implementation plan for Scenario1 MST-Simple

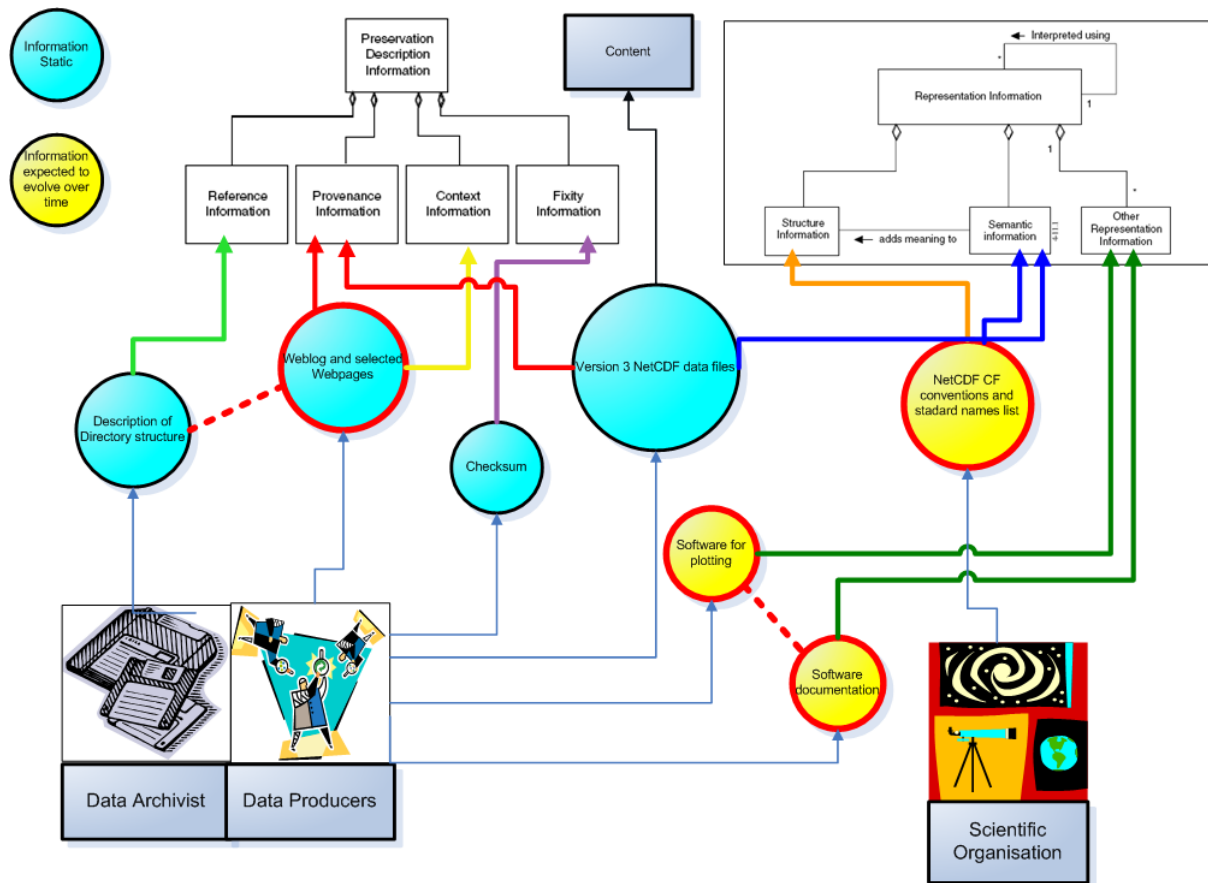
The MST Radar at Capel Dewi near Aberystwyth is the UK's most powerful and versatile wind-profiling instrument. Data can currently accessed via the British Atmospheric Data Centre. It is a 46.5 MHz pulsed Doppler radar ideally suited for studies of atmospheric winds, waves and turbulence. It is run predominantly in the ST mode (approximately 2–20 km altitude) for which MST radars are unique in their ability to give continuous measurements of the three dimensional wind vector at high resolution (typically 2–3 minutes in time and 300 m in altitude).



A user from a future designated user community should be able to extract the following information from the data for a given altitude and time

- Horizontal wind speed and direction
- Wind sheer
- Signal Velocity
- Signal Power
- Aspect
- Correlated Spectral Width

5.1.7.1.1 Preservation Information Flow for Scenario1 MST-Simple



5.1.7.1.2 Implementation points based on strategies for scenario1

MST1.1 Create meaningful reference to CDAT software <http://www2-pcmdi.llnl.gov/cdat> for reading manipulation and analysis of NetCDF files. Use POM to send notifications of the status of this software development initiative.

MST1.2 Demonstrate how the GAP manager can be used to identify NetCDF file as at risk when CDAT goes away either to a variety of technical or organisational reasons and we have been notified through POM. This can now be replaced with other Repinfo from the registry repository which we will take from the NetCDF document library at UNICAR whose longevity is not guaranteed <http://www.unidata.ucar.edu/software/netcdf/docs/>. We will use this documentation and the real life BADC user survey to create different designated community profile with the GAP manager. This will show how we can satisfy the needs of different communities of C++, Fortran, Python and Java programmers who wish to use the data.

MST1.3 CDAT supports a wide variety of file formats and analysis functions. We can demonstrate using the OAIS relationship database how CDAT is used by a variety of datasets for different reason and how the Knowledge and GAP manager can support large scale management of datasets within scientific data archives

MST1.4 Demonstrate how CASPAR works well with good data management practices and initiatives developed within communities. Describe what is good about NetCDF



standardisation and show CASPAR supports it by archiving the CF standard name list monitoring it and using POM to send notification of changes therefore supporting the semantic integrity of the data.

MST1.5 Archive the MST support website and carrying out an assessment of it constituent elements and use the Registry to repository to add basic information on HTML, Word, PDF, JPEG, PNG and PostScript to facilitate preservation of a simple static website

MST 1.6 Demonstrate how the analysis methodology is flagging up risk areas for STFC as an organisation. Carry out an investigation and review of long term web archiving option for STFC web based grey materials and project wikis.

MST1.7 Use PACK to create and add checksum to the AIP maintaining the existing directory structure

5.1.7.2 Implementation plan for Scenario2 MST-Complex

A user from a future designated user community should be able to extract the following information from the data for a given altitude and time

- Horizontal wind speed and direction
- Wind sheer
- Signal Velocity
- Signal Power
- Aspect
- Correlated Spectral Width

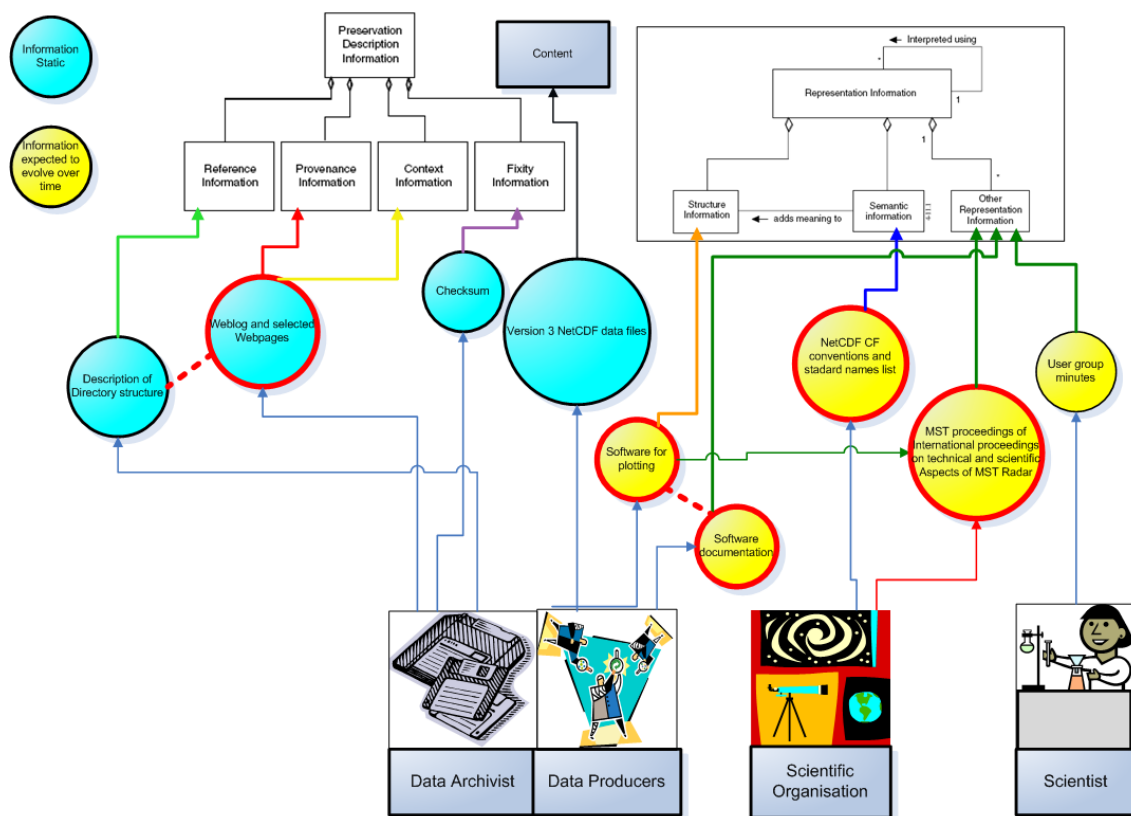
In addition future users should have access to User group notes, MST conference proceedings and peer reviewed literature published by previous data users.

MST Scenario2 has a higher level preservation objective and can be considered an extension of scenario 1 as the AIP information content is simply extended. The significance of this is that future data users will have access to important information which will help in the studying the following types of phenomena captured within the data

- Precipitation
- Convection
- Gravity Waves
- Rossby Waves
- Mesoscale and Microscale Structures
- Fallstreak Clouds
- Ozone Layering



5.1.7.2.1 Preservation Information Flow for Scenario2 MST-Complex



5.1.7.2.2 Implementation points based on strategies for scenario1

MST1.7 Review bibliography contained by website and quality of references. Carry out an investigation and review of technical reports which are used heavily at STFC but have not been generated here. Identify clear cases of reports which have correctly cited but have not need been deposited anywhere as they have no natural home and digitise for inclusion within the AIP.

MST1.8 Locate and identify the best method for referencing MST International Workshop conference proceedings held by the British Library. Identify proceeding not held by the British Library and digitise for ingestion into the AIP. Carry out an investigation and review of conference proceedings and other materials which are used heavily at STFC but have not been deposited at a national library or institution.

MST1.9 Create meaningful reference to the MST user group minutes held in the newly created CEDA institutional repository for the Nation Centre for Atmospheric studies <http://cedadocs.badc.rl.ac.uk/>. Develop an orchestration strategy for material held by this repository as it is representative of a proliferation of repositories in academia whose longevity is not guaranteed

5.1.8 Ionosonde data and the WDC

The World Data Center (WDC) system was created to archive and distribute data collected from the observational programmes of the 1957–1958 International Geophysical Year. Originally established in the United States, Europe, Russia, and Japan, the WDC system has since expanded to other countries and to new scientific disciplines. The WDC system now includes 52 Centers in 12 countries. Its holdings



include a wide range of solar, geophysical, environmental, and human dimensions data. The WDC for Solar-Terrestrial Physics based at the Rutherford Appleton laboratory holds ionospheric data comprising vertical soundings from over 300 stations, mostly from 1957 onwards, though some stations have data going back to the 1930s.

The Ionosonde is a basic tool for ionospheric research. Ionosondes are “Vertical Incidence” radars which record the time of flight of a radio signal swept through a range of frequencies (1-30MHz) and reflected from the ionised layers of the upper atmosphere (90-800km) as an “ionogram”. These results are analysed to give the variation of electron density with height up to the peak of the ionosphere. Such electron-density profiles provide most of the Information required for studies of the ionosphere and its effect on radio communications. Only a small fraction of the recorded ionograms are analysed in this way, however, because of the effort required. The traditional input to the WDC has been hourly resolution scaled data, but many stations take soundings at higher resolutions.

The WDC receives data from the many ionosonde stations around the world through a variety of means including ftp, email, CD-ROM. Data is provided in a number of formats: URSI (simple hourly resolution) and IIWG (more complex, time varying) standard formats as well as station specific “bulletins”. The WDC stored data in digital formats comprises 2.9GB of data in IIWG format and 70GB of raw MMM, SAO, ART files from Lowell digisondes. The WDC also holds about 40,000 rolls of 16/35mm film ionograms and ~10,000 monthly bulletins of scaled ionospheric data. Some of this data is already in digital form, but much, particularly the ionogram images, is not yet digitised.

- Many stations’ data is provided in IIWG or URSI format directly. This data may be automatically or manually scaled.
- A selection of European stations provide “raw” format data from Lowell digisondes, a particular make of ionosonde, as part of a COST project. This data is in a proprietary format, but Lowell provide Java based software for analysis. The WDC uses this software to manipulate this data, particularly from the CCLRC’s own Ionospheric Monitoring Groups ionosondes at Chilton, UK and Stanley, Falkland Islands. The autoscaled data from these stations is also stored in a PostgreSQL database.
- Other stations provide a small set of standard parameters in a station specific “bulletin” format which is similar to the paper bulletins traditionally produced from the 1950s onwards. The WDC has some bespoke, configurable software to extract the data from these bulletins and convert it to IIWG format.

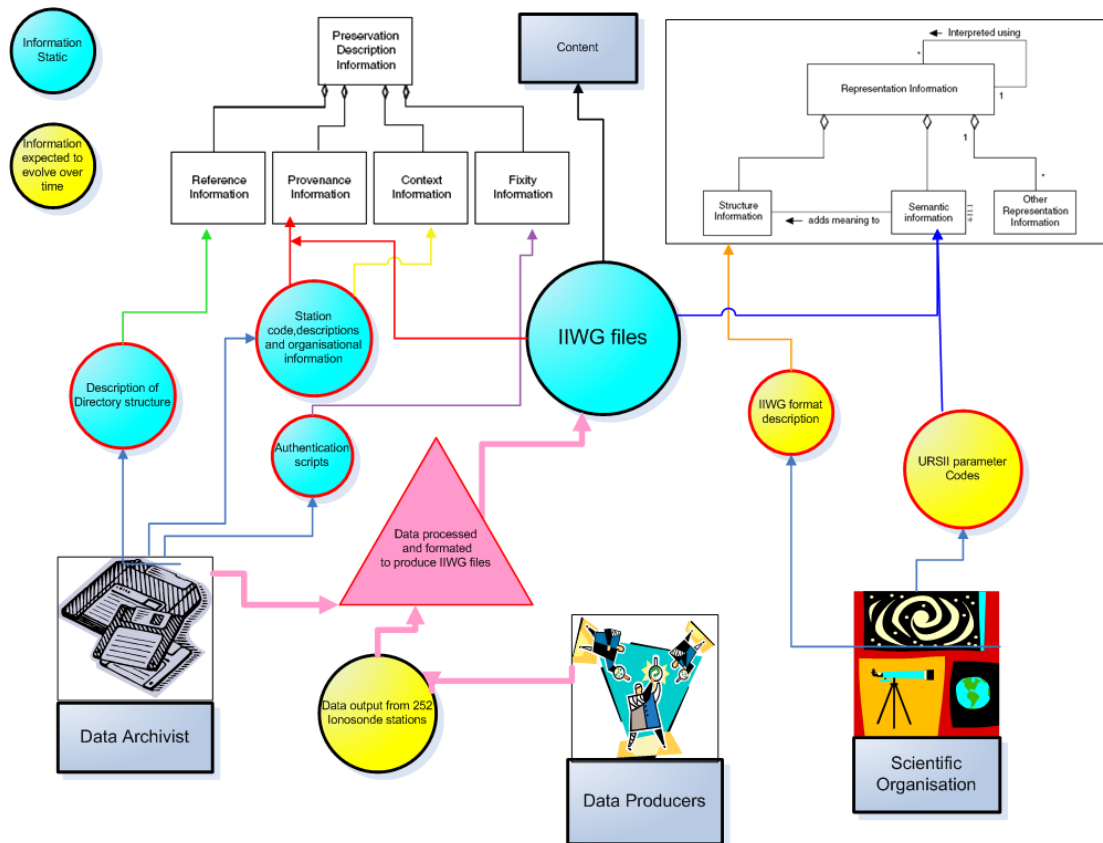
It is important to realise that this is a totally voluntary data collection and archive system. The WDCs have no control or means of enforcing a “standard” means of data processing or dissemination, though “weight” of history and ease-of-use tends to make this the preferred option.

5.1.8.1 Implementation plan for Scenario3 Ionosonde-Simple

The first preservation scenario show us again supporting and integrating with existing preservation practices of the World Data Centre, which means creating a consistent global record from 252 station by extracting a standardise set of parameters from the Ionograms produced [around the world](#). A user from a future designated community

should be able to the following fourteen standard Ionospheric parameters from the data for a given station and time. They should also be able to understand what these parameters represent. F_{min} , foE' , h_E , $foEs$, h_Es , type of Es , $fbEs$, $foF1$, $M(3000)F1$, h_F , h_F2 , $foF2$, fx , $M(3000)F2$

5.1.8.1.1 Preservation Information Flow for Scenario3 Ionosonde-Simple



5.1.8.1.2 Implementation points based on strategies for scenario3

IO1.1 Create new RepInfo based on IIWG format description removing need to understand FORTRAN as is the case with comprehending the current version

IO1.2 Create DEDSL dictionary for 14 standard parameters and add RepInfo from the Registry Repository on the XML DEDSL standard

IO1.3 Gather and model authenticity information from the current archivist for the 252 stations and the data transformation/ingest process

IO1.4 Perform CSV dump of station information from Postgres database

IO1.5 Create logical description of directory structure

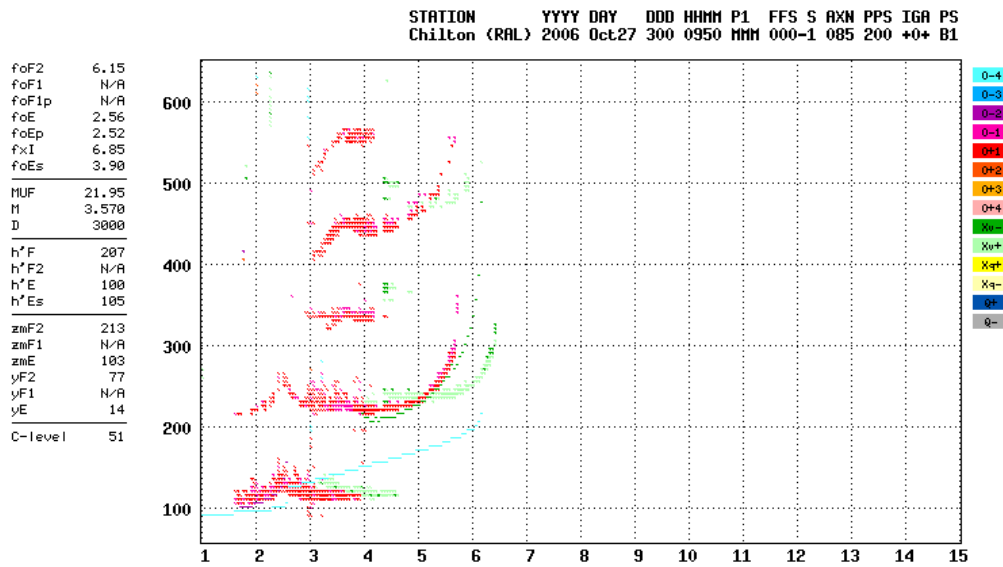
IO1.6 Use PACK to create and add checksum to the AIP maintaining the existing directory structure

5.1.8.2 Implementation plan for Scenario4 Ionosonde-Complex

The second preservation scenario for the Ionosonde can only be carried out for 7 European stations but will allow a consistent Ionogram record for the Chilton site which

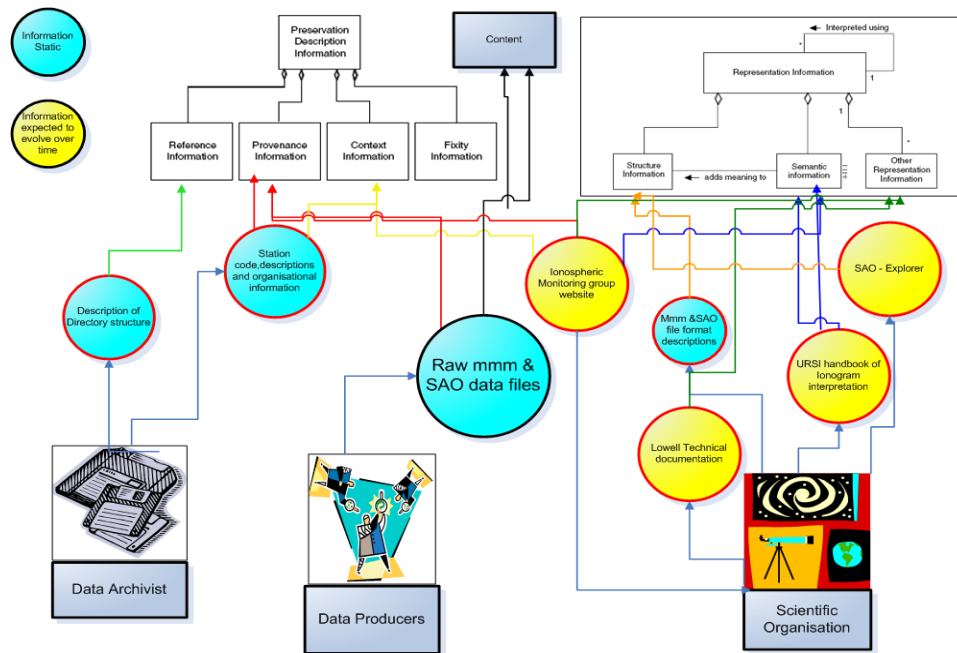


dates back to the 1920's. A user from a future designated community should be able reproduce an Ionogram from the raw mmm/sao data files and have access to the Ionospheric Monitoring groups website, the URSII handbooks of interpretation and Lowell technical documentation. Being able to preserve the Ionogram record is significant as it a much richer source of information more accurately able to convey the state of the atmosphere when correctly interpreted.



/data/ionosondes/chilton/2006/10/RL052_2006300095000.MMM / 280fx128h 50 kHz 5.0 km 2x3 / DPS-1 (052-052) S1+6 N 358.7 H

5.1.8.2.1 Preservation Information Flow for Scenario4 Ionosonde-Complex



5.1.8.2.2 Implementation points based on strategies for scenario4

IO2.1 Archive SAO explorer with RepInfo from registry repository for JAVA 5 software

IO2.2 Digitise and include URSI handbooks of interpretation in the AIP and deposit in Registry Repository for other repository users

IO2.3 Digitise and include Lowell technical documentation in the AIP and deposit in Registry Repository for other repository users

IO2.4 Archive the Ionospheric monitoring group website and carrying out an assessment of its constituent elements and use the Registry to repository to add basic information on HTML, Word, PDF, JPEG, PNG and PostScript to facilitate preservation of a simple static website

IO2.5 Review bibliography contained by website and quality of references. Carry out an investigation and review of technical reports which are used heavily at STFC but have not been generated here. Identify clear cases of reports which have correctly cited but have not need been deposited anywhere as they have no natural home and digitise for inclusion within the AIP.

IO2.6 Perform CSV dump of station information from Postgres database

IO2.7 Create logical description of directory structure

IO2.8 Use PACK to create and add checksum to the AIP maintaining the existing directory structure

IO2.9 Use the GAP manager to identify a GAP based on the demise of the JAVA virtual machine. Use POM to notify us of the gap and update the AIP with a replacement EAST description of the mmm file structure from the registry repository.



5.2 ESA

Authors: Sergio Albani (ESA), Marco Fulcoli, Fulvio Marelli (ACS)

5.2.1 Abstract

ESA plays the role of both user and infrastructure provider for the scientific data testbed; the selected dataset consists of data from GOME (Global Ozone Monitoring Experiment), a sensor on board the ESA ERS-2 (European Remote Sensing) satellite.

The complete dataset is composed by data of different levels, processors and auxiliary data needed to generate higher level data starting from raw data, documents and methods, data viewers, examples of science applications, format converters etc..

Note that the GOME dataset is just a demonstration case because similar issues involve many other Earth Observation instrument datasets.

The testbed activities have covered:

- the setup of the framework in ESA-ESRIN;
- the collection of a significant sample of a whole processing chain dataset;
- the conversion of data from the native format to the SAFE OAIS compliant format generating appropriate Representation Information, Descriptive Information and so on;
- the ingestion of the dataset in the CASPAR system;
- the coping with some long term data preservation problems by using the CASPAR components, methodology and tools.

ESA data preservation initiatives will benefit of the results of CASPAR by adopting, when applicable, technical solutions and procedures developed in the framework of this cooperative project.

5.2.2 Relation with Deliverable 4101

ESA preservation issues are summarized by a specific testbed scenario focusing on a process preservation. As GOME products are the result of different processings applied to an initial raw data, called Level0, which is sent to ground by the satellite, the need is to preserve not only the data but also the processors/algorithms (and related representation information) needed to generate on demand higher level products (e.g. Level1, Level2, etc.). These test cases cope with some of the scenarios described in deliverable 4101, such as changes in hardware and software (change of hardware, operative system or compilers/libraries/drivers that could affect the ability to run the Data Management System, the GOME Data Processors or the format/auxiliary data converters).

5.2.3 Testbed Description

5.2.3.1 Background and purpose

GOME Level1C data are fully calibrated products obtained by the processing of Level1B data (raw signals plus calibration data). Applying different calibrations it is possible to obtain different versions of Level1C data (note that a Level1B has a size of



about 15 Mb while the corresponding Level1C versions vary from 800Mb to several GB, this last case if all the calibration algorithms are applied).

Then, to have an affordable preservation process, only Level1B data and the processor to obtain Level 1C data starting from Level 1B can be ingested and archived. This implies that when a user needs a Level 1C data, he must be able to retrieve both the corresponding Level 1B data and all that is needed – processor, manuals, etc. – to configure the process and generate it.

The chance to preserve - through CASPAR - the ability to process Level 1B data in order to generate Level 1C data is the core part of the ESA testbed. It implies the preservation of data, software and all those info needed to use the software and to apply it to data.

5.2.3.2 The Testbed Phases

The ESA testbed can be summarized in the following phases.

Ingestion

- The data producer generates a Level 1C product by processing a Level 1B product. The processor has been set to apply all the calibration algorithms.
- The data producer ingests into CASPAR the whole Level 1B product, the processor (if not already ingested) and a proxy object representing the Level 1C product, including all the info on how to generate it from its correspondent Level 1B.

Browsing

- Two different CASPAR users, with a different knowledge base, are browsing the archive searching for a specific Level1C product. They will retrieve in CASPAR the requested data and the appropriate Representation Information needed to have a full data understanding (depending on their knowledge).
- The users need to re-create the Level 1C product: thus they ask for a L1C and they are returned with the corresponding Level 1B plus the processor and all the Representation Information that they need to perform the processing.

System update needed

- An event affecting the possibility to run the L1B->L1C processor happens (e.g. a new OS version has been released); this event generates the need to access the source code of the processor, to deal with the new situation and to ingest a new version of the processor into the system.
- At the same time all the Representation Information needed to understand the processing must be updated. CASPAR will update all the links between processor and data, in order to keep track of changes and guarantee the user the possibility to continue to perform the L1B product processing.

5.2.4 Testbed CASPAR components

The Testbed is built on top on the following CASPAR components

- Pack Manager;



- PDS;
- FindingAids;
- Registry;
- GapManager;
- Orchestration Manager;
- DAMS;
- (facultative) RepInfo Toolbox and Virtualization Assistant.



5.2.5 Testbed development time schedule

According to the Testbed «phases» described, the testbed will be developed adopting the following schedule.

Phase 0

Objective(s)	Testbed set-up		
Tasks	<ul style="list-style-type: none"> • Choice of an adequate infrastructure • Definition of Descriptive and Representation Information for GOME Level1B/C data • Definition of Descriptive and Representation information for the L1B->L1C processing • Definition of Designated communities profiles • User interface design 		
Start Date		End Date	November 08

Phase 1

Objective(s)	Ingestion and browsing		
Tasks	<ul style="list-style-type: none"> • Integration with Pack Manager for AIP creation • Integration with Finding Aids for data search and retrieval • Integration with PDS for data storage • Integration with DAMS for user profiling • Basic integration with Registry and Gap Manager for RepInfo management • (facultative) Basic use of Virtualization assistant and RepInfo toolbox to generate new RepInfo • Graphical user interface development 		
Start Date	November 08	End Date	February 09

Phase 2

Objective(s)	Processor update needed		
Tasks	<ul style="list-style-type: none"> • Full integration with Gap Manager and Registry • Integration with Orchestration Manager • Integration with Virtualization Assistant 		
Start Date	February 09	End Date	March 09

Phase 3

Objective(s)	Test, validation and refinement		
Tasks	1. Final Caspar Testbed Integration		



	2. Collect user feedback 3. Analysis user feedback 4. Make recommendations		
Start Date	March 09	End Date	May 09

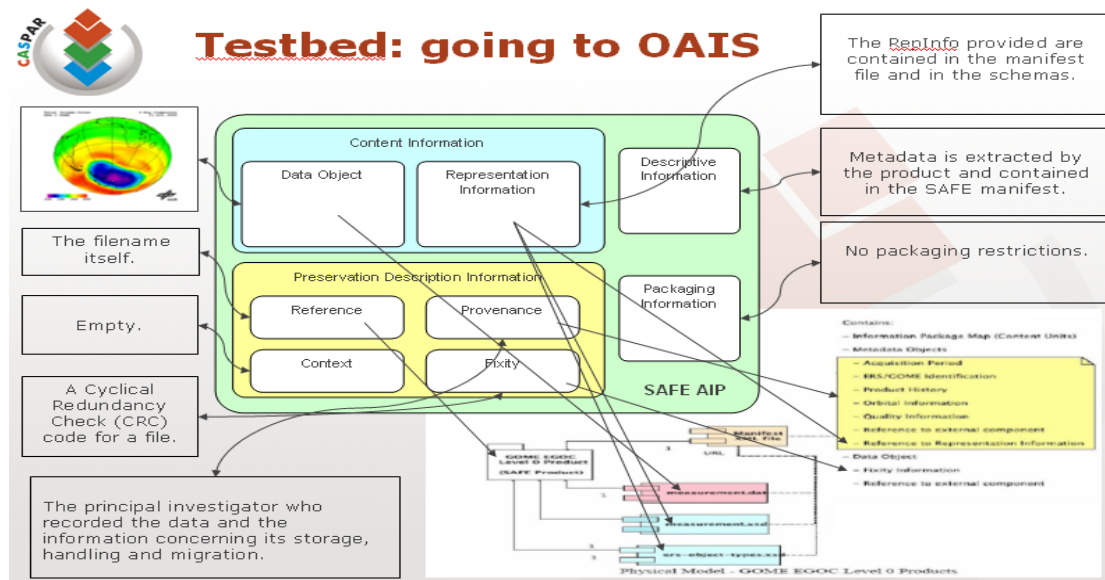
5.2.6 GOME dataset and designated communities

5.2.6.1.1 GOME Dataset Definition

The GOME Dataset to be ingested is comprehensive of the following files.

- GOME Level 1B products (YYYY/MM/DD/*.lv1)
 - PSD.pdf: *Level 1 and Level 2 Product Specification Document*
 - license.doc: *License for the GOME data products on FTP and CD*
 - disclaimer.pdf: *Disclaimer for GOME Level 1 and Level 2 data product summarizing the status of the current GDP data quality*
 - ERS-Product.pdf
 - ProductSpecification.pdf
- Processors (precompiled versions of the Level 1 extraction software)
 - user_manual.pdf: *GDP Level 1 and Level 2 Extraction Software Manual*
 - howtouse_101.doc: *Brief explanation on how to use the software*
- 'C' files with the Level 1 extraction software (source code).
 - readme_1st.doc: *Summary of files and documents*
 - release_101.doc: *GDP Level 0->1 Processing Release Notes*
 - programmer's manual

The following picture maps the ESA GOME L0 data to the OAIS standard:



In a similar way, AIP for the L1 products composed by a manifest file (with PDI) containing files with RepInfo, DescInfo and data have been produced.

User Profiles have been created to show the dependency between known modules and different Representation Information that will be returned to the user during a browsing session.

5.2.7 ESA Scientific Testbed Scenarios

This chapter is aimed to give a brief description of the scenarios composing the ESA Testbed.

5.2.7.1 Data Ingestion

Step	Functionality	Caspar elements	Additional info
1	Data creation		
2	Gome L1B SIP creation	Registry	Search for RepInfo
3	Gome L1C SIP creation	Registry	Search for RepInfo
4	Processor creation	SIP Registry	Search for RepInfo
5	AIP creations	PACK	
6	AIP storage into CASPAR	Find, PDS	Insert data into pds



5.2.7.2 Data Search and Retrieval

In this phase two different profiles are supposed to search and access Level 1C data; this to show how profiling impacts on RepInfo visualisation.

Step	Functionality	Caspar elements	Description
1	User n°1 Login	DAMS	First User profiling
2	Data search and retrieval	FIND, PDS	Formalisation of CPID search criteria
3	Level 1C RepInfo browsing	PACK, GAP, REG	The user asks to see the result of his query. He wants to browse some generic Representation Information related to the Level 1C. Depending on his profile the user gets different levels of RepInfo.
4	User n°2 Login	DAMS	Second User profiling
5	Data search and retrieval	FIND, PDS	Formalisation of CPID search criteria
6	Level 1C RepInfo browsing	GAP, REG	The user asks to see the result of his query. He wants to browse some generic Representation Information related to the Level 1C. Depending on his profile the user gets different levels of RepInfo.

5.2.7.3 Level 1C data creation

This part of the scenario has been isolated from the previous section to highlight the operation of re-creation of a Level 1C data, starting from its ancestors.

Step	Functionality	Caspar elements	Description
7	Level 1C ancestors visualisation	Same as point 6	The user is able to point to the Level1B data object and to the related processor needed to generate the Level 1C data. Level 1C contains all needed RepInfo by which the user can know how to perform the processing.
8	Level 1C Data Obj Creation		The user has the DIP for Level 1B file and correspondent



			processor. He creates locally a new Level1C Data Object.
--	--	--	--

5.2.7.4 Software change

This part of the scenario is based on the hypothesis that some software must be produced/updated to maintain the ability to process data from Level 1B to Level1C. This could be a new processor for a new release of the operative system or a new processor with a new licence.

Step	Functionality	Caspar Elements	Description
1	External change		
2	Experts notification	POM	
3	Retrieve of the Source Code of the processor, new processor coding and compiling		
4	Creation and ingestion of the new Processor AIP	PACK, PDS, FIND, Registry	
5	RepInfo and DescInfo fixing	Find, Registry and Gap Manager	

5.2.7.5 New Data browsing

This part of the scenario is a replication of part 2; it is used to demonstrate that the new processor is available and properly working.