



Project no. 033572

## CASPAR

**C**ultural, **A**rtistic and **S**cientific knowledge for **P**reservation, **A**ccess and **R**etrieval

**Instrument:** Information Society Technologies

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

# D4101 USER REQUIREMENTS AND SCENARIO SPECIFICATIONS



---

Document identifier:	<b>CASPAR-D4101-SCEN-0101-1_0</b>
Date:	<b>21-Dec-2006</b>
Work package:	<b>4100</b>
Partners:	<b>All Partners</b>
WP Lead Partner:	<b>CCLRC</b>
Document status	<b>Delivered</b>

---

**Abstract:** This document provides the first instalment of User Requirements and Scenario Specifications based on analyses of detailed questionnaires about specific datasets



<b>Delivery Type</b>	Report
<b>Author(s)</b>	CASPAR Consortium
<b>Approval</b>	David Giaretta
<b>Summary</b>	This document provides the first instalment of User Requirements and Scenario Specifications based on analyses of detailed questionnaires about specific datasets.
<b>Keyword List</b>	User requirements, validation, testbeds, scenarios
<b>Availability</b>	<input checked="" type="checkbox"/> PUBLIC

### Document Status Sheet

Issue	Date	Comment	Author
0_0	29 Sept 2006	Initial draft bringing together contributions from all testbeds	Simon Lambert (editor)
0_1	15 Oct 2006	Revised draft based on comments from all partners	David Giaretta (editor)
1_0	21 Dec 2006	Final release	David Giaretta (editor)





### Project information

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

### 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@cec.eu.int

### Project Co-ordinator: David Giaretta

Address:	CCLRC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	<a href="mailto:d.i.giaretta@rl.ac.uk">d.i.giaretta@rl.ac.uk</a>





## CONTENT

<b>DOCUMENT STATUS SHEET .....</b>	<b>2</b>
<b>PROJECT INFORMATION .....</b>	<b>3</b>
<b>1 INTRODUCTION.....</b>	<b>7</b>
1.1 PURPOSE OF THIS DOCUMENT.....	7
1.2 DOCUMENT STRUCTURE .....	7
1.3 SCOPE .....	7
1.4 CASPAR OBJECTIVES .....	8
<b>2 THE APPROACH TO REQUIREMENTS ACQUISITION .....</b>	<b>10</b>
2.1 QUESTIONNAIRES.....	10
2.1.1 <i>Pre-questionnaire</i> .....	10
2.1.2 <i>Structure of the full questionnaire</i> .....	11
2.1.3 <i>Candidate projects/datasets</i> .....	11
2.1.4 <i>Pre-questionnaires and prioritisation</i> .....	12
2.1.5 <i>Use of full questionnaire</i> .....	12
2.1.5.1 Science examples .....	12
2.1.5.2 Performing arts examples.....	12
2.1.5.3 Cultural heritage examples.....	13
2.2 ANALYSIS APPROACH .....	14
2.2.1 <i>Preservation issue identification</i> .....	14
2.2.2 <i>Special issues</i> .....	15
2.2.3 <i>Preservation scenario creation</i> .....	15
2.2.3.1 Hardware and software changes.....	15
2.2.3.2 Environment changes.....	16
2.2.3.3 Simulation of changes in the designated community .....	16
2.2.3.4 Additional considerations.....	17
<b>3 THE SCIENCE DOMAIN .....</b>	<b>18</b>
3.1 WORLD DATA CENTRE IONOSONDE DATA (CCLRC).....	18
3.1.1 <i>Introduction</i> .....	18
3.1.2 <i>Preservation significance</i> .....	20
3.1.3 <i>Preservation issues</i> .....	21
3.1.4 <i>Preservation scenarios</i> .....	27
3.1.4.1 Changes in hardware and software.....	27
3.1.4.2 Changes in environment.....	28
3.1.4.3 Changes in designated community.....	29
3.2 EISCAT DATA (CCLRC).....	30
3.2.1 <i>Introduction</i> .....	30
3.2.2 <i>Preservation significance</i> .....	31
3.2.3 <i>Preservation issues</i> .....	32
3.2.4 <i>Preservation scenarios</i> .....	36
3.2.4.1 Changes in hardware and software.....	36
3.2.4.2 Changes in environment.....	36
3.2.4.3 Changes in designated community.....	36
3.3 MESOSPHERE-STRATOSPHERE-TROPOSPHERE (MST) RADAR DATA (CCLRC).....	37
3.3.1 <i>Introduction</i> .....	37
3.3.2 <i>Preservation significance</i> .....	39
3.3.3 <i>Preservation issues</i> .....	40
3.3.4 <i>Preservation scenarios</i> .....	45
3.3.4.1 Changes in hardware and software.....	45
3.3.4.2 Changes in environment.....	45
3.3.4.3 Changes in designated community.....	46
3.4 EUROPEAN SPACE AGENCY EARTH OBSERVATION DATA TESTBED.....	47
3.4.1 <i>Introduction</i> .....	47
3.4.1.1 Background.....	47
3.4.2 <i>Preservation significance</i> .....	49
3.4.2.1 ESA test-bed .....	51





3.4.3	<i>Preservation issues</i>	52
3.4.4	<i>Preservation scenarios</i>	57
3.4.4.1	Changes in hardware and software	57
3.4.4.2	Changes in environment	57
3.4.4.3	Changes in designated community	57
<b>4</b>	<b>THE ARTS DOMAIN</b>	<b>58</b>
4.1	THE IRCAM TESTBED	58
4.1.1	<i>Introduction</i>	58
4.1.2	<i>Preservation significance</i>	59
4.1.3	<i>Preservation issues</i>	61
4.1.4	<i>Preservation scenarios</i>	64
4.1.4.1	Changes in hardware and software	64
4.1.4.2	Changes in environment	65
4.1.4.3	Changes in designated community	65
4.2	THE INA TESTBED	66
4.2.1	<i>Introduction</i>	66
4.2.1.1	The Acousmatic Music production scenario	66
4.2.2	<i>Preservation issues</i>	69
4.2.3	<i>Preservation scenarios</i>	76
4.2.3.1	Changes in hardware and software	76
4.2.3.2	Changes in environment	76
4.2.3.3	Changes in designated community	76
4.3	THE UNIVERSITY OF LEEDS TESTBED	77
4.3.1	<i>Introduction</i>	77
4.3.2	<i>Preservation Scope</i>	77
4.3.3	<i>Preservation issues</i>	79
4.3.4	<i>Preservation scenarios</i>	82
4.3.4.1	Changes in hardware and software	82
4.3.4.2	Changes in environment	83
4.3.4.3	Changes in designated community	84
4.4	THE CIANT TESTBED	85
4.4.1	<i>Introduction</i>	85
4.4.2	<i>Preservation issues</i>	86
<b>5</b>	<b>THE CULTURAL HERITAGE DOMAIN</b>	<b>93</b>
5.1	THE WORLD HERITAGE SITE TESTBED (UNESCO)	93
5.1.1	<i>Introduction</i>	93
5.1.2	<i>Preservation significance</i>	95
5.1.3	<i>Preservation issues</i>	98
5.1.4	<i>Preservation scenarios</i>	101
5.1.4.1	Changes in hardware and software	101
5.1.4.2	Changes in environment	101
5.1.4.3	Changes in designated community	102
<b>6</b>	<b>COMMON REQUIREMENTS</b>	<b>103</b>
6.1	CHANGES IN HARDWARE AND SOFTWARE	103
6.1.1	<i>Changes in storage technologies</i>	103
6.2	CHANGES IN ENVIRONMENT	103
6.2.1	<i>Changes in legal framework</i>	105
6.3	CHANGES IN DESIGNATED COMMUNITY	105
<b>7</b>	<b>CONCLUSIONS</b>	<b>106</b>
	<b>REFERENCES</b>	<b>107</b>
<b>A1</b>	<b>OTHER COMMON REQUIREMENTS FROM THE WARWICK WORKSHOP</b>	<b>108</b>
A1.1	COMMON RESEARCH ISSUES IDENTIFIED ACROSS ALL THREE DISCUSSION GROUPS	108
A1.2	SPECIFIC RESEARCH TOPICS	108
A1.3	POLICY AND INFRASTRUCTURE DEVELOPMENT	109
<b>A2</b>	<b>THE PRE-QUESTIONNAIRE</b>	<b>111</b>





---

**A3 FULL QUESTIONNAIRE ..... 112**





## 1 INTRODUCTION

### 1.1 PURPOSE OF THIS DOCUMENT

This document provides the first instalment of requirements for the CASPAR components. Further instalments will be produced at various stages through the life of the project.

The steps involved have been to:

- (1) examine a number of datasets in detail, identifying, as far as possible, all types of information implicitly or explicitly used by knowledgeable users to extract usable information from bit sequences.
- (2) identify a number of issues, requirements and testbed scenarios covering, as far as possible, all aspects of changes which might affect the preservability of the information encoded in bit sequences.

### 1.2 DOCUMENT STRUCTURE

The requirements must be viewed in the terms of the scope of CASPAR, which is discussed next, and also the overall goals and objectives of the project, which, for convenience, are repeated, in section 1.4, from the Description of Work.

Section 2 describes the way in which we have approached gathering the requirements and scenarios. Given the wide range of disciplines it was important to elicit details in a structured way which would allow intercomparison of needs. The guiding principles for this have been the concepts from the OAIS Reference Model.

Sections 3, 4 and 5 have examples from, respectively, science, performing arts and cultural heritage. Each example is analysed in a similar way. Each example covers a range of different digital data types. We also give some indication of the relevance of the specific examples to the broader world of digital data.

In section 6 the common requirements (but not an exhaustive set) are collected and some guiding principles are proposed. These are supplemented by the requirements taken from the Warwick workshop [WARWICK-1] in Appendix A1, which are closely related to CASPAR, as is the Research Programme proposed by the Task Force on Permanent Access to the Records of Science [TFPA].

### 1.3 SCOPE

Digital preservation involves legal, social, financial as well as technical issues. All are important but clearly CASPAR must be focussed on the technical solutions, although we would hope that there would be financial impacts because the availability of more effective techniques. The increased ability to share preservation efforts, which the CASPAR system will allow, will reduce the costs of digital preservation for data archives. CASPAR may also be able to provide some insight into the legal, social and financial aspects of preservation but is unlikely to provide solutions in these areas.

The OAIS Reference Model provides the high level architecture for the technical issues which CASPAR must address. However while the Reference Model provides fundamental concepts such as *Representation Information* and *Designated Community*, in order to put produce a real-world implementation we must confront real-world requirements. We must also ask fundamental questions such as *What does Representation Information include?*, *Is it practical to capture enough?*, *Can a Designated Community be adequately defined?* and many others, in order to test the usefulness of OAIS itself.

There are many strategies which may be adopted to preserve digitally encoded information. The OAIS Reference Model identifies four primary digital migration types namely (1) refreshment and (2) replication, both of which do not change the bit sequences, (3) repackaging which changes the Packaging Information and (4) transformation, which may





change the bit sequence of both the Packaging and the Content Information. OAIS also discusses preservation strategies which depend upon keeping the bit sequences unchanged and preserving the access to the information. Access which relies on emulation implies an unchanged bit sequence. Underlying these strategies is the need to maintain the understandability and usability of Digital Objects (bit sequences), by means of the use of Representation Information.

Considering *transformation* in particular, it is probably safe to assume that those strategies which are human resource intensive will **not** continue indefinitely. Also the question arises as to what to transform the bit sequence to at each stage, bearing in mind the need for further preservation efforts and the need to ensure that the bit sequences are usable by the Designated Community. Experience needs to be gained about the implications of the use of particular formats or families of formats to inform the choice. Furthermore the actual encoding of the information is only a small part of the Representation Information required, in particular the amount of relevant semantics (explicit, implicit and tacit) may be very large and much will be independent of the coding, i.e. independent of the format.

**For these reasons CASPAR will focus on tools and techniques to preserve the ability to convert a given bit sequence into usable information. We will, over the life of the project, take a very wide range of example datasets to identify the capabilities required. On the basis of the experience gained from this we will then be in a position to advise on preservation strategies and produce relevant tools and infrastructure components.**

It is worth noting that benefits can accrue from preservation activities to contemporary users as well as to future generations. We say this because preservation efforts are designed to keep digitally encoded information understandable and usable by future generations, despite its being *unfamiliar* to them; but the same is true for that same information to many, if not most, contemporary users. Their unifying idea is therefore that of being able to allow users to understand *unfamiliar data*, whether created yesterday or many decades ago.

## 1.4 CASPAR OBJECTIVES

From the CASPAR Description of Work, the objectives of the project can be stated as follows:

The **CASPAR** challenge is to achieve four main goals that can be stated as follows:

**Goal 1:** build a pioneering preservation environment, based on a full use of the OAIS Reference Model<sup>1</sup> and building in the latest developments in knowledge technologies

**Goal 2:** demonstrate its ability to handle the preservation of the digital resources of many user communities

**Goal 3:** advance the current state of the art in digital preservation

**Goal 4:** development of technological solutions supporting the emergence of an offer of systems and services for preservation of digital resources

Expanding these goals into more specific objectives, CASPAR will:

1. Implement, extend and validate the OAIS reference model.
2. Enhance the techniques for capturing Representation Information and other preservation related information for content objects.
3. Design virtualisation services supporting the preservation of digital resources over the long term, despite changes in the underlying computing (hardware and software) and storage systems, and the Designated Communities.
4. Integrate as standard features of CASPAR, digital rights management, authentication and accreditation.





5. Research more sophisticated access and use methods of preserved digital resources including intuitive query and browsing mechanisms.
6. Develop case studies demonstrating the validity of the CASPAR approach to the preservation of digital resources across different user communities and assessing the conditions for a successful replication.
7. Actively contribute to the relevant standardisation activities in areas addressed by CASPAR.
8. Raise awareness about the critical importance of the preservation of digital resources among the relevant user-communities and facilitate the emergence of a more diverse offer of systems and services for preservation of digital resources.

Validation of these objectives is to be via a number of Measurable Objectives, addressing preservation aspects including

- a sound theoretical basis and in particular alignment with the OAIS Reference Model,
- "accelerated lifetime" tests involving hardware, software and the knowledge base of the Designated Community,
- an increase in the trustworthiness of archives using CASPAR.





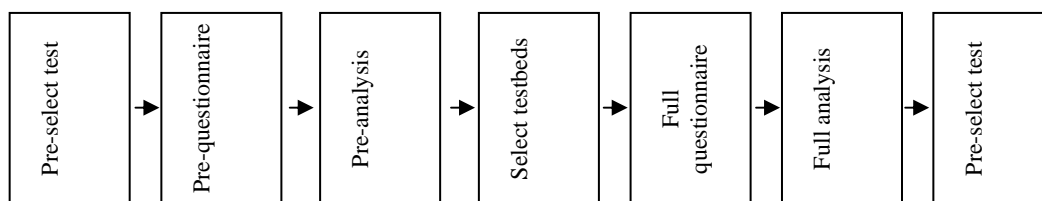
## 2 THE APPROACH TO REQUIREMENTS ACQUISITION

The methodology adopted was developed by taking the best aspects of the InterPARES [InterPARES-1] and ERPANET [ERPANET-1] questionnaire methodologies, but structuring the resulting questionnaire much more strongly in line with OAIS and focussing on specific datasets. However, in order to prioritise the datasets, specific selection phases were introduced based on a “pre-questionnaire”.

The resulting methodology involves five stages:

1. Prepare pre-questionnaire and full questionnaire, based on OAIS, the CASPAR validation metrics, and additional sections designed to elucidate digital rights management (DRM), authenticity and provenance.
2. Identify candidate repositories/projects/datasets in the three domains (science, cultural heritage and performing arts).
3. Obtain input from pre-questionnaires and produce prioritised list.
4. Interviewers then obtained information to answer the full questionnaires from a prioritised list of candidates, based on the pre-questionnaires.
5. Subsequent analysis identified appropriate scenarios for testbeds and requirements for components and the framework.

For each repository the following diagram shows the process flow.



### 2.1 QUESTIONNAIRES

The questionnaires were structured to be aligned with the draft high level CASPAR architecture and components, but with sufficient flexibility to capture the preservation plans which may already have been in place for each dataset.

#### 2.1.1 Pre-questionnaire

The pre-questionnaire was a very short version of the full questionnaire and designed to be completed by data managers by themselves if necessary. The aim of the pre-questionnaire was to be able to select a number of datasets to investigate in more depth.

Each repository was characterised at the outset by the following basic features.

1. Holdings: overview of the type of data held, and a list of data sets.
2. Data Set: A description of the digitally encoded information to be preserved, from the bit level to the knowledge it conveys to its user community. We do not at this stage need very detailed descriptions. In addition we need a brief description of
  - a. access restrictions
  - b. what information/behaviour the data encodes





- c. how the data is stored
  - d. how the required data is located and retrieved (including DRM and Legal issues)
  - e. what additional data, equipment or knowledge is employed to extract required information/behaviour from the data.
3. Data Producer: A brief description of the group, individual or institution that produced the data set.
  4. User Community: A description of the current user community and the characteristics of the designated community for whom this data might be preserved.
  5. Current preservation plans.

### 2.1.2 Structure of the full questionnaire

The questionnaire is structured following the CASPAR architecture, components and framework. Thus it covers:

For each dataset, more details of:

1. production: description of the way in which the information is captured or created
2. current use:
  - finding aids
  - software used to access the digital encodings
  - software/mechanisms to use/perform the encoded information

and, in alignment with the CASPAR high level architecture:

- ingestion into the repository
- access control
- knowledge/behaviour encoded
- domain specific virtual objects e.g. sound recordings, moving images, Earth observation images, Solar Terrestrial Physics datasets – these can be made from:
  - generic virtual objects e.g. images, tables, sequences, etc plus simple values
  - binary encodings of the information
  - storage mechanisms

The questionnaire is available on the CASPAR web site ( <http://www.casparpreserves.eu> ) and also as an appendix to this document. Completed pre-questionnaires and full questionnaires are available as separate documents.

### 2.1.3 Candidate projects/datasets

Each domain – science, culture and arts – produced a list of candidate projects/repositories, where possible specified down to individual datasets. The list of candidates aimed to cover a variety of repositories, data types, producers, access controls, significance of the information etc. This list will be extended throughout the project.





### **2.1.4 Pre-questionnaires and prioritisation**

The prioritisation criteria were allowed to be different between the domains. One essential criterion was to ensure that there was a good variety in each group of, say, ten in each discipline, in order to ensure that there was, at a minimum, an adequate variety. A small number (at least two) of cases were then identified for each domain as suitable to taking into testbed scenarios, and a great deal more information was gathered for these cases.

### **2.1.5 Use of full questionnaire**

The full questionnaire was written to be understandable by the interviewers, who have a common understanding of the questions, and enable them to obtain a minimum depth of detail for each case. The interviewers talked to a variety of people, as appropriate, involved with each repository in order to gather information to complete the questionnaire.

Although the questionnaire was designed to be applied to a single dataset, it has proved convenient to combine a number of related datasets. A number of points will be applicable to all these datasets, with additional issues related to specific datasets identified separately.

#### **2.1.5.1 Science examples**

The CCLRC questionnaires were completed in collaboration with the UK Digital Curation Centre (<http://www.dcc.ac.uk>) CCLRC's Ionosonde testbed is characterised by the preservation needs of data accumulated over a long period of time, and from geographically distributed sources. In essence the same quantities are always being measured, but with different instruments, processing software, etc. The processing chain from raw data is of key importance, especially the knowledge embedded in it, for example about the characteristics of particular ionosondes. A proprietary file format, whose structure is not publicly available, is used by one of the leading Ionosonde manufacturers. It would be desirable to allow for annotations, for example of peculiarities of particular measurements, so that data can be correctly interpreted in the future.

CCLRC's EISCAT testbed shows some of the same preservation issues. In particular, the processing chain applied to the data is of key importance. Multiple independent analyses can be made of the same raw data, and the significance of these should be preserved. A similar requirement for annotation arises, for example to indicate whether filtering was applied to the raw data. Some of the data processing has a dependence on external reference models, outside the control of the archive, that are updated over time. It would be desirable to capture and preserve the intent of particular experiments and the links to other (non-EISCAT) measurements made as part of the same experiment.

CCLRC's MST testbed, like the Ionosonde testbed, holds a long-term record of atmospheric data. Once again the processing chain is important, a particular issue being the progressive enhancements of the signal processing techniques represented by updated versions of the data processing software. As for all science testbeds, data authenticity and integrity arises from trust in the organisation's staff and reliability of the underlying operating systems and computer hardware.

The ESA Earth observation testbed is characterised by a processing chain on the raw data, and the emergence and adaptation to a new standard archive format. Supplementary data includes data on satellite orbits and calibration. High-level knowledge covers the purpose of observation campaigns.

#### **2.1.5.2 Performing arts examples**

The IRCAM testbeds in electroacoustic music and allied fields, by contrast with the science testbeds, throw up issues of the purpose of retrieval of archived data, whether for viewing, re-performance or analysis and criticism. There are diverse types of materials required to fully represent a performance - texts, active modules, technical documentation, etc. The





intentions of the creator of the work are important, as is the judgement of authenticity of future performances. To permit re-performance, the archive must allow the retrieval of everything needed in a usable form.

INA's testbed in acousmatic music production shows strong similarities to IRCAM's. Again it is concerned with preserving the musical integrity of the work. There is a dependence on technology such as sound sequencers, which are in constant evolution, making it difficult to freeze particular states for preservation. There is varied supplementary information of relevance, including not only performance directives but also recordings of past performances and external material such as essays and newspaper articles.

The University of Leeds testbed in Interactive Multimedia Performing Arts, like the science testbeds, makes the processing chain of key importance. The chain links the creative input (e.g. gestures of the performer) to the output work. The use of specialised software and hardware in interactive multimedia systems will complicate the problem, as any replacement of components may cause a loss to the integrity of a performance. A particular issue is that there are many different file formats currently used for 3D motion data, usually specific to the applications they work with.

CIANT's testbed, the AMANT archive, for video art encoded in Real Video, allows annotation by curators. The aim is to provide access to a set of distributed archives similar to AMANT maintained by other institutes.

### **2.1.5.3 Cultural heritage examples**

UNESCO's testbed concerns World Heritage Sites. This is of great importance because the information has legal status, and also because the sites themselves change and deteriorate over time, so the archive provides a record of their state. The testbed will focus not only on the storage, retrieval and preservation of the basic data, but also on preserving the associated knowledge such as how the data was acquired and the software needed to interpret it. A requirement is to be able to merge data and extract different models such as 3D representations. New data formats (such as map data) might be defined outside the archive and the archive must be able to cope with this kind of evolution of practice.





## 2.2 ANALYSIS APPROACH

### 2.2.1 Preservation issue identification

In general modelling terms we have, as described in the Description of Work [DoW], a preservation model guided by OAIS, which suggests the main components needed. A detailed Conceptual Model will be presented in Work Package 1200.

In normal modelling terms what we need next are high level Use Cases - these are scenarios which show some aspect of added value - in this case preservation. From these Use Cases requirements are then extracted.

In order to guide the production of scenarios, and in particular to have adequate coverage of the 'preservation space' we use the following table to identify issues:

<b>CASPAR element</b>	<b>Summary of Current situation of testbed</b> (Likely source of information in questionnaire is shown below)	<b>Preservation issues arising from</b>
		<ul style="list-style-type: none"> <li>• <b>changes in hardware and environment (software, legal, social etc) and Designated Community</b></li> <li>• <b>loss of sources of Information (including loss of host archive)</b></li> </ul>
Ingest	Section 4 of questionnaire	
Preservation Description Information	Addressing issues of Provenance, Fixity, Context and Reference	
Representation Information	Parts of sections 2, 3 and 9 of questionnaire	
Annotation		
Packaging		
Description Information		
Access	Section 5 of questionnaire	
Access control, including DRM	Sections 6 and 14 of questionnaire	
Higher-level knowledge	Section 2 of questionnaire and the separate FORTH questionnaire (treated as section 15 in CCLRC's questionnaires)	
Virtualisation and representation information	Sections 10, 11, 7 and 8 of questionnaire	
Storage and storage virtualisation	Section 12 of questionnaire	
Preservation orchestration	Some aspects of section 9 of questionnaire	
Authenticity	Some aspects of section 4 of questionnaire and others	





This method of identifying issues focuses on extracting requirements for the CASPAR components and framework, and also testing the adequacy of the overall CASPAR architecture.

Where the analysis reveals that we need further information/clarification on our data background or its preservation requirements, then we go back to the repository with further questions.

### **2.2.2 Special issues**

During discussions with the various stakeholders, it is clear that many have in mind improvements to the current system; such improvements are not within the remit of CASPAR. However where these improvements reflect an underlying preservation requirement, for example to capture the knowledge held only by specific individuals, who can be asked for facts and clarifications right now – but not in the future, then these ideas should be captured, and the corresponding preservation issues noted.

### **2.2.3 Preservation scenario creation**

It is clearly important to generate a wide range of scenarios in order to touch upon as many preservation issues as possible. In order to guide this process we propose a classification of potential changes, based on themes highlighted by the OAIS Reference Model. Within each of these broad areas we try to describe, in general terms, the range we should consider when constructing the scenarios. In this way we have a checklist against which we can measure the coverage of our testbed work.

#### **2.2.3.1 Hardware and software changes**

The range of computer CPU hardware is becoming more diverse, when taking into account new technologies such as RISC, Cell processors, embedded systems and also updates to existing CPU instructions sets (affecting binary code compatibility). The increase in popularity of virtual architectures such as Java and .NET only adds to the problem of the diversity and future compatibility of ‘binary application code’.

There are some specialised pieces of hardware which may be involved in scenarios and careful consideration must be given to functionality and maintenance.

Similarly, if one takes into account the many different versions of Linux, BSDs, Solaris, Windows, embedded operating systems, OpenVMS etc the diversity of operating system environments has increased over time. As for running application code, it is not always possible to reliably run a Linux application on the many different flavours of Linux (even with the same CPU type) due to differences in library and compiler versions, which means that different Linux flavours can all be considered to be different operating systems. Similar things can be said for the other operating systems.

There is a much greater range of application software which we will have to cover in some way. A given software application will be written in a specific language or languages (e.g. Java, Fortran, C), which implies the need for the appropriate compilers, possibly more than one for a given language. The application is also linked, dynamically or statically, with other things such as system libraries and device drivers. The latter will frequently change as hardware such as storage devices, human interface devices and network devices.

Scenarios can include:

- Change of operating system for processing s/w
- What actions are needed as a result?

The nature of the change may include the following:





- Possible decay of storage medium
- Obsolete hardware or software (this will encompass obsolete file formats)
- Change in availability of software or documentation (copyright issues)

### **2.2.3.2 Environment changes**

Environment changes include:

- organisational environment e.g. organisation staff
- legal environment, including digital rights

Copyright restrictions on data, software, hardware and supporting information evolves over time, new rights may be created and others may expire. The result of this evolution is the need to ingest or release access to the previously identified materials, information or data. Some form of monitoring of the required copyrighted materials and owning institutions is required to facilitate this.

It should be noted that copyright changes do not occur at the same time for all related information for all end users i.e. the copyright expire on a score/data from scientific experiment but supporting text books /journal articles or instructional manuals is still under copyright. This creates a common requirement across all data sets that access be dependant upon the unique combination of end user type and information unit retrieved.

Changes in law e.g. freedom of information or data protection have a very broad effect. All archives are potentially subject to evolving restrictions be they at a governmental or organisational level (i.e. between an archive and a user community). The ability to apply restrictions based on individual information units and specific types of user within the designated user community is required.

### **2.2.3.3 Simulation of changes in the designated community**

Special consideration must be given to simulating changes in the Designated Community. The aim in the testbed would be to verify the ability to ensure that data continues to be independently usable and understandable by the Designated Community as its knowledge base changes. CASPAR should be able to supply tools which enable enough Representation Information to be created and updated.

We propose a general set of scenarios for any particular dataset as follows:

- Produce a number (e.g. 12) of questions which someone who understands that dataset should be able to answer – the current set of data users should be able to produce this list.
- Choose sets of users each defined by knowledge bases increasingly different from the current users. At least some of the sets of users should be able to answer the questions satisfactorily. For example if the dataset concerns Ionosondes then one could have
  - Current users: very knowledgeable about ionosondes
  - Ionospheric physicists
  - Atmospheric physicists
  - Physicists
  - Scientists





- Non-scientists

It is not reasonable to demand that all sets of users be able to provide satisfactory answers - if only because of the limited time available for performing the tests, but it should be possible to get somewhat down the list. Special consideration must also be given to Performing Arts, where special skills are necessary such as playing the required instrument.

#### **2.2.3.4 Additional considerations**

- Can the change be detected by the CASPAR orchestration process? If so then how?, If not then why not?
- Are the implications of changes worked out by the CASPAR process?
- Does the CASPAR process identify and support the actions required?



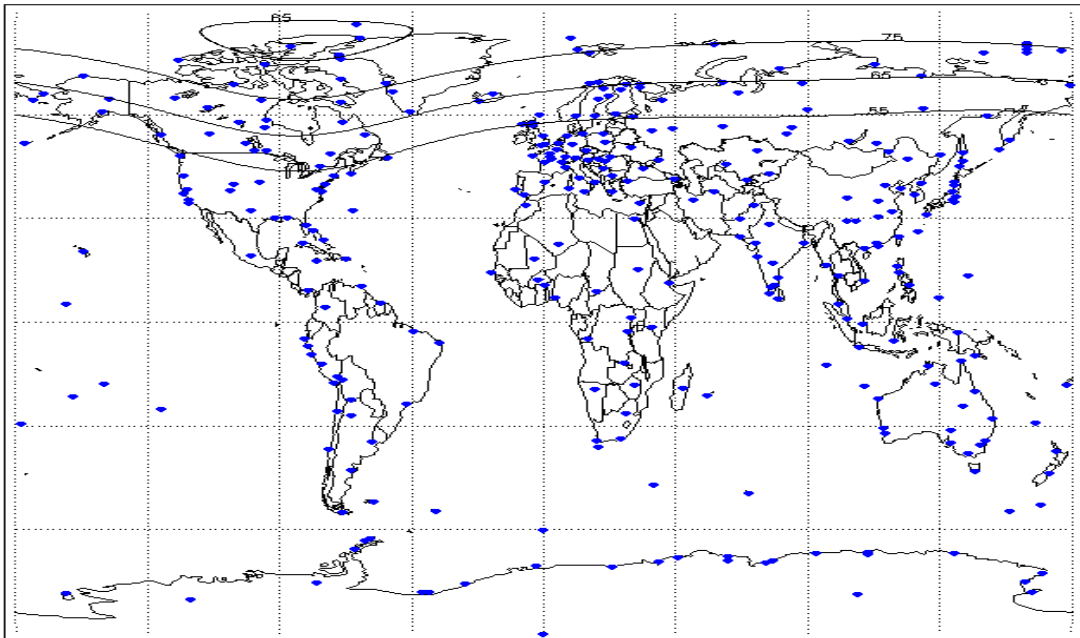
## 3 THE SCIENCE DOMAIN

### 3.1 WORLD DATA CENTRE IONOSONDE DATA (CCLRC)

#### 3.1.1 Introduction

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.



**Illustration 1: Worldwide distribution of ionosonde stations**

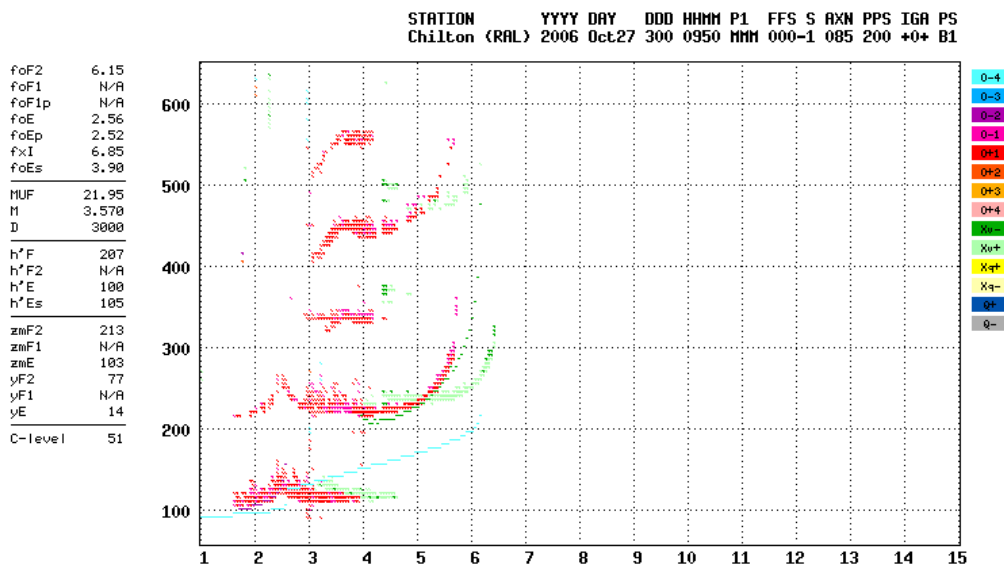
The WDC receives data from the many ionosonde stations around the world through a variety of means including ftp, email, CD-ROM. Data is provided in a number of formats: URSI (simple hourly resolution) and IIWG (more complex, time varying) standard formats as well as station specific “bulletins”. The 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.



/data/ionosondes/chilton/2006/10/RL052\_2006300095000.MM / 280fx128h 50 kHz 5.0 km 2x3 / DPS-1 (052-052) 51.6 N 356.7 W

## Illustration 2: Sample ionogram from a Lowell digisonde

Most ionosondes are provided by a small number of commercial companies which may also provide proprietary analysis software.

Recreating the actual plasma density profile from ionogram data is an important use of ionosonde data. Such a procedure is known as real or true height analysis.





The time of flight from each echo (frequency sweep) in an ionogram, gives some indication of the height at which the radio wave was reflected. This cannot be taken as the true-height of the layer due to the effect of any ionisation in the path of the wave.

In order to obtain true height values, the whole ray path must be reconstructed and this requires assumptions to be made about the electron concentration along the ray path. The assumptions which allow one to do this are embedded in community knowledge and software which have evolved over time.

As a testbed this collection of data allows us to explore data resulting from same type observation using similar types of instrument over long period of time. It will permit us to examine the issues surrounding data originating from geographically diverse locations, operated by different types of organisation, using different models of instrument, employing different modes of operation whilst additionally being subject to different software and manual interpretive processing. It will also allow us to inspect the evolution and preservation requirements of the knowledge base that surrounds a relatively mature area of scientific investigation.

### 3.1.2 Preservation significance

**Importance:** The primary interests of the organizations who co-operate to provide this global network of observational data fall into four different fields. The data, being environmental monitoring is also impossible to reproduce and may have future relevance not currently seen. As seen by its recent use in global climate change studies.

- Those primarily concerned with earth environment studies the data allows for long term global monitoring and mapping
- Those interested in the exact form of the ionosphere at a specified time, e.g. for comparison with a rocket or satellite data or for studying time variation in events.
- Those primarily concerned with radio propagation problems and communications research, both surface and space.

For those involved in geophysical studies the data also allow one to examine various types of geophysical phenomena. The current WDC user community comprises approximately 1000 users annually with over 2 million individual data accesses.

**Uniqueness of data and holding:** The WDC based at Rutherford Appleton Laboratories is part of a larger international WDC system which is continually evolving due to scientific, technical and economic factors. When the WDCs were originally set up in 1957, multiple centres were deemed advisable to guard against catastrophic loss of data, and for the convenience of data providers and users. According to the WDC the system is currently healthy and viable with most centres maintaining their funding, though not without struggle.





### 3.1.3 Preservation issues

CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and
Ingest	<p>The ingestion process can be quite complex and relies on the personal knowledge of the archive staff and internal documentation as this is not always a fully automated process in addition to the use of software developed in-house.</p> <p>Data arrives from the global network of ionosondes by a number of methods as well as being in a number of different formats. This data may require additional processing so it is in the correct format for deposit i.e. IIWG or extracted ionospheric parameters.</p> <p>On occasion the data does not automatically arrive and the archive staff personal knowledge is relied upon to make note of this and contact the appropriate organisation to obtain the appropriate data if possible. The quality of the archive also relies on the archive staff maintaining the ingest from the global network.</p> <p>There are additionally a number of processes which occur post the initial ingest which extract parameters, update and populate the postgresQL databases and manage the directory structures.</p>	<p>Complexity of ingest process, including role played by staff of the archive in detecting and acting on anomalies.</p>
Preservation Description	<b>Provenance:</b> Source implicitly linked to file location which	<b>Provenance:</b> Should be formalised including source of





Information	<p>identifies station of origin. There is additional station, instrument and organisational information held within a postgresQL database. There is also much additional information held in the community relating to the make, model and mode of operation of the ionosondes, information on scaling processes and organisational information which is not currently within the archive.</p> <p><b>Fixity:</b> reliant on trust and correctness of operating system</p> <p><b>Reference:</b> full Unix path and filename within archive, not externally accessible.</p> <p><b>Context:</b> WDC staff and scientist knowledge.</p>	<p>data, method of transfer. Table of station information is vital.</p> <p>For processed data the processing steps should be detailed in Provenance. Relevant projects include [PROV-EU] and [PROV-MYG]</p> <p><b>Fixity :</b> need to investigate techniques for confirming fixity e.g. digests. Also the security of any particular digest algorithm over time.</p> <p><b>Reference:</b> Persistent Identifiers</p> <p><b>Context:</b> see [CCLRC-01]</p>
Representation Information	<p>File extension indicates format</p> <ul style="list-style-type: none"> <li>• MMM = proprietary raw data format from Lowell digisonde. An EAST [EAST] description is now available</li> <li>• SAO = scaled data – full format description available</li> <li>• ART = proprietary format of automatically scaled data (full format description not released by Lowell)</li> <li>• IIWG = ionospheric parameters– full format description available</li> </ul> <p>Some parameter data only exists within the Postges databases.</p> <p>Data dictionaries available for all paramaters contained with databases and IIWG file. There is also much additional information held by the community. Additional information is contained within URSI handbooks for ionospheric information and other key texts we have scanned into PDF for the purposed of this project. There is additionally material contained within standard texts, journals and websites on ionospheric science which are consulted regularly by</p>	<p>Formal language descriptions of the data formats and data dictionaries would allow independence from some of the access software.</p> <p>The ART format is proprietary, used by proprietary software, and may need continuing support, or else the details of the format must be obtained and maybe kept in a 'dark archive'.</p> <p>The POLAN software is open source and requires source code plus a FORTRAN compiler and UNIX operating system.</p> <p>There is a fundamental issue about the amount of Representation Information: does the INFORMATION object from the RAW data include the processed results? If so then this would imply the whole processing stream constitutes Representation Information. However this would imply that all potential processing schemes would also be Representation Information. This extreme interpretation seems unsustainable. Some middle way needs to be</p>





	users of the data.	found.
Annotation (could be regarded as a special type of Rep. Info.)	Not explicitly captured by the WDC but there may be several separate analyses of a particular piece of raw data.	Some kind of explicit annotation system would be relevant to this data. In addition there are many scientific theories which the ionospheric monitoring group have not had the time and resources to fully investigate and therefore be recorded in journals. There are also other unexpected occurrences such as a large industrial fire or volcanic eruption which would have an impact on the behaviour of the ionosphere the ability to annotate affected data is highly desirable
Packaging	Simple files in directories. File name includes code for source station and date. No additional packaging	Raises the question of the reliability of filename and directory structure, for example if one transfers a file to another system with different file structure. Name and date information exist as values within some of the file formats. Definition of the AIP is needed - perhaps [XFDU]
Descriptive Information	Finding Aids: Descriptive Information about the file is placed in database used by finding aid and by the archive. Details of station location etc stored in separate table.	
Access	Access via web based finding aids using PERL scripts to identify data file from postgresQL database. There are many special processing options. For example: Data is currently accessed via the web by specifying in the type of information <ul style="list-style-type: none"> <li>• Instrument</li> <li>• Data Availability Listings</li> <li>• IIWG Parameters</li> </ul>	Access and processing software currently require PERL, POSTGRES, FORTRAN (C-wrapper), C and C++. Does this software need to be preserved? Issue: the finding aid table can be re-constructed from data as long as the directory structure is maintained, but needs additional Information about location of station which obtained the data. Should the local "filename" actually be regarded as the full path?





	<ul style="list-style-type: none"> <li>• Prompt Data (Data availability listings, Autoscaled Parameters, Autoscaled Parameter Plot, Autoscaled POLAN height profiles, Autoscaled NHPC height profiles, Autoscaled Parameters Sec Style, Autoscaled File Download, Manual Parameter Data)</li> <li>• Raw Data Files</li> <li>• Recent Ionogram products (Ionogram, F-plot, Autoscaled Parameters, Manually checked parameters, True height(POLAN) profile on Autoscaled data, True height(POLAN) profiles on manually scaled data)</li> </ul> <p>In addition specifying the type of information station and time are also required. Various programs are also used to extract reformat and in some cases carry on further processing of the data In order to create the required digital object.</p>	
Access control, including DRM	<p>Read access open to all for data files although registration is required and a record of accesses is kept. No copyright restrictions are imposed. Important to note is that copyright restrictions currently act as a barrier to ingesting critical representation information. See also section 14 of [CCLRC-01].</p> <p>There are issues with respect to non-CCLRC documentation and manuals.</p> <p>Write access limited to system operations.</p>	Loss of access to manuals would seriously affect the usage of the data. These manuals may only be available in paper form and only in limited numbers, and may be copyright protected. This raises the question of what we can rely on in the long term and how that list is made explicit.
Higher-level knowledge	<p>See section 2 of [CCLRC-01]. There is a great deal of associated information</p> <p>No formal ontologies exist although the data dictionary may considered to be one which defines all the IIWG parameters and many parameters within the other file formats relationships between</p>	<p>Loss of information on external web sites would affect usage of the data. Related issues:</p> <ul style="list-style-type: none"> <li>• simple web site</li> <li>• content managed site may be especially difficult</li> </ul>





	<p>the parameters have not been defined. The data dictionary carries some instrument, organisational and other provenance information.</p>	<ul style="list-style-type: none"> <li>• depth of cross-linking to other sites also difficult</li> </ul> <p>Related issue of what paper documentation can be relied upon e.g. journals and other cross-references. Manuals were referred to under "Access control" above.</p> <p>Loss of expert staff would make use of data more difficult. Some information is held only by a specific individual e.g. hand written notes, email or just human memory.</p> <p>Software such as POLAN, a “true-height” profiling program, is required to process the data - the algorithm behind it is important to preserve. Software SAO_trace is required to extract auto/manually scaled ionogram trace information from SAO files fro procoessing by POLAN software.</p> <p>Processing algorithm of Lowell software is not public. Software SAO_pars is required to extract the NHCP profile from the SAO files.</p>
<p>Virtualisation and representation information</p>	<p>Some data is represented as simple tables:</p> <ul style="list-style-type: none"> <li>• Instrument Records</li> <li>• Data Availability Listings</li> <li>• IIWG Parameters</li> <li>• File Availability</li> <li>• Autoscaled Parameters</li> <li>• Polan Height Profiles</li> <li>• Autoscaled NHPC height profiles</li> <li>• Characteristics</li> </ul>	<p>Tabular description of the raw and processed data would allow at least basic access.</p> <p>The high level parameters are relatively simple elements stored in a database. POSTGRES software would have to be available unless the information is exported to a simpler format. Even if POSTGRES is available, separate definitions of the columns would have to be available.</p>





	<ul style="list-style-type: none"> <li>• Contour (electron density profile)</li> <li>• Info (station constants etc)</li> </ul>	
Storage and storage virtualisation	Simple files, with automated backup systems to separate tape data store. POSTGRES database used, and also off-site copies kept for some data in other WDC sites.	
Preservation orchestration	Co-ordination between set of World Data Centres provides some orchestration.	This co-ordination could be related to the CASPAR-type orchestration.
Authenticity	Relies on trust in WDC staff, WDC systems and also source systems	Digests etc could be introduced - see Fixity above. Fundamental question of authenticity - including the various copy processes which happen throughout the life of a data file. A significant Issue Is the issue of authenticity of database elements.





### 3.1.4 Preservation scenarios

#### 3.1.4.1 Changes in hardware and software

##### **Scenario 1: Change of operating system affecting ability to run WDC bespoke software.**

The WDC has a range of internally produced software, Perl scripts; C and FORTRAN programs, for the manipulation and processing of standard IIWG and SAO data files.

- IIWG manipulation – combine-iiwg, display-iiwg, medians-iiwg, merge-iiwg, month-iiwg, split-iiwg, verify-iiwg – Standard C programs for appropriate manipulation of IIWG files. Requires ANSI C compiler for recompilation on new system.

Documentation on IIWG format freely available therefore reproduction of this software is feasible if necessary. Requires preservation of IIWG format documentation.

- SAO extraction – sao\_find, sao\_itec, sao\_pars, sao\_split, sao\_trace – Standard C programs for discovery and extraction of particular sections of SAO files. Requires ANSI C compiler for recompilation on new system.

Documentation on SAO format available therefore reproduction of this software is feasible if necessary. Requires preservation of SAO format documentation – multiple versions!

Perl scripts – There are various perl scripts in use that pull these software components together for data ingestion and dissemination via the WDC web site. Requires perl interpreter and WDC internal documentation web site.

##### **Scenario 2: Change of operating system affecting ability to run POLAN program (FORTRAN).**

Need to preserve ability to create POLAN profiles. The WDC website currently allows the generation of POLAN profiles automatically from the ionogram trace extracted from SAO files (see section4 CASPAR\_01). The ability to run the POLAN analysis program on these values needs to be preserved. Either the ability to run the Titheridge POLAN analysis program to produce these profiles needs to be preserved (requires standard FORTRAN-77 compiler) or the archive needs to provide sufficient documentation to recreate the data processing algorithms within new software (Titheridge algorithms in UAG-93 and scientific literature).

Not all ionosonde data corresponding to parameters is held In SAO & MMM files. Some exists in paper format or in different digital file format information at different Institutions. An extension to the basic requirement would be the inclusion of documentation and the preservation of software programs that would allow trace information from these sources to have POLAN analysis performed on it (see UAG report - 93 Ionogram analysis with generalised program POLAN for further details)

##### **Scenario 3: Change of operating system affecting ability to run View to Gif program (C++)**

Need to preserve ability to create ionograms from raw data files either by preserving viewtogif (developed at Lowell). Either software and the ability to run it must be preserved or sufficient documentation to allow for the development of automated extraction in the future. Requires full documentation of MMM format from Lowell,

##### **Scenario 4: New operating system no longer capable of running postgresQL**

Relationship between station Information and data directories needs to be preserved so all modes of access to data is preserved (see section 5 CASPAR\_01) when new database





system is queried. Requires PostgreSQL or data migration to similar ANSI/ISO SQL database management system.

### 3.1.4.2 Changes in environment

#### 3.1.4.2.1 Rights Management changes

##### **Scenario 5: Lowell no longer provides or supports SAO explorer**

Need to preserve the ability to scale ionospheric parameters and ionogram trace from SAO and MMM files (see section 7 CASPAR\_01). Either the SAO explorer program (Java) needs to be preserved along with the ability to run the program with accompanying user documentation.

##### **Scenario 6: Changes in copyright ownership and legal restrictions on materials supporting knowledge base.**

A core requirement for the preservation of the knowledge extract from and usability of the data set are copyrighted or otherwise restricted materials which include:

- Journal Articles
- Bibliographies
- Books (standard texts)
- Websites
- Software
- Technical Manuals
- Copyright restriction and ability to deal copyright issues is major reason for much of the material not being added to the archive.

The Archive would benefit from guidance as to which framework they should be operating under for the types of materials mentioned above and implications for cross border supply and International Intellectual Property rights .

An alert service advising of key changes affecting such types of materials may be of use. As would automated alert to the clearance of copyright on specified (by archive manager) material such as key journal or texts so they be digitised and added to the archive or released from a dark archive.

#### 3.1.4.2.2 Organisation and personnel changes

##### **Scenario 7: Retirement of key personnel affecting knowledge base**

The standard texts contain the mature well established scientific theory. There is a very significant preservation risk with this material as the majority of it is out of print with no record of numbers/holders of copies or organisational responsibility for preservation. The number of texts one would wish to store within the archive is dependant on the depth and breadth of knowledge one is seeking to preserve so it is difficult to provide and estimate of this.

Current end users have become dependent on subscription products such as web of science <http://scientific.thomson.com/products/wos/> to locate journal material. This constitutes a significant preservation risk and our ability to preserve such listings along with relevant subject indexing is something that requires further investigation.

The result is a need to create an unbiased bibliography in consultation with key people in the field of Ionospheric Science (see section 9 CCLRC\_01)

There are lot of theories within the community which have not had sufficient research done to make it into journal literature. The annotation of these theories to the appropriate data





would be welcomed as would identification of events affecting the ionosphere, comments on scaling quality or behaviour of the ionosonde affecting quality of observation.

### **Scenario 8: Collapse of organisation supporting knowledge base**

Organisational monitoring is significant so previously withheld information may be requested from an organisation which may have lost its funding become bankrupt etc. An archive would benefit from research into the information support provided by the listed organization. As such information would enhance the level and quality of knowledge that can be extracted from an archive if it could be preserved by incorporation into the archive

Organisational support materials are again a high preservation risk. Mechanisms and strategies for digitising storing and cataloguing a wide range of materials which originate from diverse groups would be highly desirable

The content of external website should be monitored for changes and any relevant material preserved within the archive if deemed to be at risk

#### **3.1.4.3 Changes in designated community**

### **Scenario 9: Change in user community**

As science evolves the reason why a scientist may wish to use this type of observational data. It is extremely difficult to attempt to anticipate how a user community will change. However connecting this data into larger ontology which evolves over time for atmospheric science would assist the discovery and use of this data.

Use proposed methodology from section 2.2.3.3.





## 3.2 EISCAT DATA (CCLRC)

### 3.2.1 Introduction

The European Incoherent Scatter (EIScat) association was founded in 1975 by the research councils of Norway, Sweden, Finland, France, Germany and the United Kingdom to build and operate a research radar system in the auroral regions of Scandinavia; Japan joined the association in 1996. The technique uses high power radio waves at UHF and VHF frequencies that are weakly scattered by the ionosphere. Measuring the spectra of the returned signal allows routine measurement of Electron Density, Ion Velocity and Electron and Ion temperatures from about 80km to over 1000km with height resolution down to a few hundred meters. There are a number of other incoherent scatter systems in the world (e.g. Sondre Stromfjord, Millstone Hill, Aereceibo) but the EISCAT UHF system is unique in that it has two remote receivers at Kiruna and Sodankylä that enable full three dimensional plasma velocities to be measured.

The raw signal from the ionosphere is stochastic in nature; auto-correlation is used as a signal detection system and the resulting functions are integrated for typically 1-5 seconds. If in this period any hard target (e.g. a satellite) enters the beam then the signal is swamped by this echo. There is also an ionospheric heater on the site and this can be used to conduct experiments on the plasma in the ionosphere.

The raw data is typically further integrated to a level of 1 to 5 minutes before being passed through some analysis program that extracts the parameters above. Under ideal circumstances this analysis could also produce parameters for the ion composition and the ion-neutral collision frequency as well as those given above; often however the analysis relies on the help of a model to reduce the degrees of freedom in the fitting process in order to produce a result.

Data can be further processed by folding in a magnetic field model (e.g., IGRF) and a neutral atmosphere model (e.g. Alcadye) to produce values for Joule heating and ionospheric conductivity.

EISCAT also operates a VHF transceiver in Tromsø and a UHF system located in Longyearbyen on Svalbard.

Taking data from Tromsø and the two remote stations allows a full three dimensional velocity vector to be computed. Under very dynamic conditions the theory behind the analysis breaks down, a so called “non-maxwellian” plasma, and features of this can be seen in the spectra from the remote sites where for instance the temperature of the plasma depends on the direction from which one looks.

Operating time on the radar is divided into two classes: a “common programme” that is intended to form a synoptic set to enable long term studies and is shared equally by all members of the association and a “special programme” where each associate runs the radar using programmes created by the associate with the privilege of exclusive use for one year. Much of the special programme time is used in multi-associate collaborations where again those collaborators share the exclusive use.

The radar is also used to make observations of the solar corona using the technique of inter-planetary scintillation.

In normal operations a simple analysis of the data is carried out in near-real time. The results of this analysis are then copied into an archive in Kiruna. The raw data is sometimes analysed/reanalysed subsequently to the operation and this data may also be copied to the archive in Kiruna. The UK group mirrors data from this archive but also add extra analyses done in the UK by staff members at CCLRC (RAL) or by selected “expert” users from the university community. A limited amount of analysed common programme data is shared with the world community though a programme known as “cedar” and a database referred to as “madrigal”.





The specific data holding selected for the CASPAR testbed is the CCLRC mirror of the archive at Kiruna.

### 3.2.2 Preservation significance

The incoherent scatter technique provides the widest ranging measurements of the ionosphere over an extensive height range. The mainland UHF system is unique in the world having the ability to produce true three dimensional velocity data and to make observations of non-maxwellian plasma effects.

Limited amounts of the data are available without restriction, most data sets are limited to members of the association.

The raw data from the radar is both large and almost impossible to understand without the comprehensive analysis system. The analysed results are much more accessible to a non-specialist but lack a clear audit trail to ensure their veracity.

A major topic of study is the link between events seen by spacecraft and their effect on the earths ionosphere (most of the interaction between the solar wind and the earths atmosphere happens in this region).

The radar is often used in conjunction with optical experiments to help in the understanding of auroral physics.

Used with a tomographic receiving chain to determine ground truth and a vertical profile for the tomographic chain.





### 3.2.3 Preservation issues

CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and
Ingest	<p>The raw and processed data from the receivers all go into the archive held at Kiruna. This data, for the UK and common programmes, is downloaded to CCLRC in the UK from the Kiruna web server. Scripts (in <i>tcl</i>) are used to compare the data (raw and processed) with what has already been obtained.</p> <p>The raw data is stored as MATLAB files. These are processed using the <i>guisdap</i> software into result files (<i>rslt</i>), a process involving integration and repeated running of the analysis programme. The analysis is a semi-automatic process with some human intervention. The processing chain is not recorded.</p> <p>The archive directory structure has one top-level directory per year, and then files with a naming convention (see under representation information below).</p> <p>There are two additional relevant programs. For the tristatic system, <i>velcom</i> (written in C, freely available) does velocity combination. It takes data from the three stations and converts to 3D vector velocity (result files → result file). The conductivity program takes the Tromsø results file and magnetic field model and generates more data for the results file.</p>	<p><b>changes in hardware, software and Designated Community</b></p> <p><b>loss of sources of Information (including loss of host archive)</b></p> <p>The programs used for analysis will be subject to the usual preservation risks.</p> <p>The processing chain should be recorded (also a provenance issue).</p>





<p>Preservation Description Information</p>	<p><b>Provenance:</b> There is information that is not recorded about the integration: how it was done, especially whether a filter was used to detect anomalies such as hard targets; the integration strategy; the analysis strategy (e.g. ion composition model); the version of <i>guisdap</i>; plus basic information about who performed it, where and when.</p> <p>As scientists can perform their own analyses and upload them to the CCLRC archive, it is possible that there might be multiple analyses of the same data.</p> <p>The archive does not identify collaborative programmes other than as 'SP'.</p> <p><b>Fixity:</b> Reliant on trust in people and correctness of operating system.</p> <p><b>Reference:</b> local file names</p> <p><b>Context:</b> If the analysis is done at CCLRC, then an 8-day run goes into one file identified under the start date only. By contrast, if it is done online, then it is split across day boundaries.</p>	<p><b>Provenance:</b> There is a need to record the processing chain.</p> <p>There is a need to record more information about the integration process, and to provide a trail of the origin of analysed files.</p> <p><b>Fixity:</b> No specific issues.</p> <p><b>Reference:</b></p> <p><b>Context:</b> Individual files may relate to others (e.g. split across days) and this information is at risk of being obscured.</p>
<p>Representation Information</p>	<p>The file naming convention has some flexibility, e.g. to enable handling of different versions, though there is no meaning in these variations.</p> <p>All files have filename extension <i>.rslt</i>. A separate file is needed to turn this data into meaningful quantities (e.g. electron temperature in degrees Kelvin). This coding is done by CEDAR, using a documentation file available on the web. However the document is not helpful for processing so there is a simpler way, with a file <i>ncar.var</i>. The text file is read through the API reading the <i>var</i> file, and converts the <i>ncar</i> to something useable. There are Matlab, IDL and C APIs.</p>	<p>The file naming convention embodies knowledge that might be lost.</p> <p>The mapping from the contents of the results files to meaningful quantities is key. The CEDAR description and <i>ncar.var</i> file should be subject to preservation.</p> <p>The result file format requires supplementary information for full preservation, such as uncertainty estimates (if negative, meaning that the value was taken from a model).</p>
<p>Annotation (could be</p>	<p>None</p>	<p>Annotation should allow for: presence of hard objects</p>





regarded as a special type of Rep. Info.)		<p>(e.g. satellite); coherent echo; plasma line contamination; equipment failure (currently in paper log book).</p> <p>Data users should have ability to annotate data and instrument performance. Also to indicate that they have published a paper relating to a particular data set.</p> <p>Many experiments are done in collaboration with satellites or optical cameras it would be good to record this somewhere.</p>
Packaging	None	
Descriptive Information	<p><b>Finding Aids:</b> The catalogue of raw data contains the start and end date of each record, the site, some information on the experiment, the size of the data file, location in Atlas tape store.</p>	
Access	<p>Straightforward through website. The data is stored in a Unix file system (rather than a database) and a perl script is used for access.</p>	<p>There is no reliable way of matching the data to the programme under which it was acquired. At present this is a haphazard procedure relying on the knowledge of the experts.</p> <p>In the long-term, it will probably be desirable to search at a higher level than simply by date, e.g. by type of programme or pattern of antenna movement.</p>
Access control, including DRM	<p>Access to the Kiruna archive is restricted by IP address.</p> <p>There is a distinction between data from ‘common programmes’ and ‘special programmes’ (only those associates participating are allowed to access this data). However there is no information stored in the archive on what those programmes are. It is possible to reconcile the programmes with the planned schedule, information on which is available on the EISCAT web pages.</p> <p>Data more than one year old is publicly available to all EISCAT</p>	<p>Presumably such IP based access restrictions will change. Information available on the web should be captured – may involve an associated database at the remote web site.</p>





	associates. A small fraction is released publicly.	
Higher-level knowledge	<p>Some numbers in the data (such as altitude) are certain, others are subject to uncertainties.</p> <p>At some altitudes it is necessary to take parameters from models. The transition could be controlled by the operator.</p> <p>The scan pattern of the antenna could also be high-level knowledge.</p> <p>There is no high-level knowledge recorded about the design/intention of the experiment. The Swedish site has information on what was requested to be run but there is no direct link to what was actually run.</p>	<p>Undergraduate physics is assumed.</p> <p>There is material on the EISCAT websites and CEDAR. The Open Radar Consortium has relevant material.</p> <p>The standard text is Rishbeth, <i>Introduction to Ionospheric Physics</i>. This raises the issue of availability of printed books, and the relationship to size of print run i.e. presumably it matters whether there are tens, hundreds or thousands of copies available.</p> <p>A knowledge of plasma theory and when it breaks down are required to fully understand the data.</p>
Virtualisation and representation information	The 'digital objects' produced are typically colour contour plots or 'fan plots'.	
Storage and storage virtualisation	Storage organised by date is limited in its ability to capture the experimental intent.	
Preservation orchestration	Not implemented at present. Apart from basic backups and the existence of multiple copies, there is little action currently being taken with a view to long-term preservation.	
Authenticity	The usual reliance on trust in staff and systems.	





### 3.2.4 Preservation scenarios

#### 3.2.4.1 Changes in hardware and software

##### Scenario 1: Unavailability of generic software tools

There is a dependency on MATLAB for the file formats of raw and processed data. It is possible that MATLAB licences might become too expensive, or that for other reasons the dependency on MATLAB has to be broken. This requires an abstract description (EAST) of the file formats.

##### Scenario 2: Unavailability of specialised software

Due to changes in operating system (*inter alia*), specialised software such as *guisdap* and the velocity combination and conductivity programs might cease to be available. The implications would be that the link between the raw and processed data would be broken, as it would not be possible to perform processing with this software. Either the software should be preserved in runnable form, or the algorithms should be represented at high level to allow reimplementing of the same functionality.

#### 3.2.4.2 Changes in environment

##### Scenario 3: Dependence on external models

There is some dependence on external (and changing models) that are not under the control of the archive. The International Geophysical Reference Field is used by the conductivity program, reading its coefficients to use in processing. This reference model is updated every five years. It is therefore necessary to preserve the model so that in future it will be known what version was used.

#### 3.2.4.3 Changes in designated community

##### Scenario 5: Change in user community

Use proposed methodology from section 2.2.3.3

##### Scenario 6: Termination of organisation supporting archive

If the organisation supporting the archive were to cease operation, or cease supporting the archive, the files might well be stored somewhere but the ability to understand them would decay. For example, the significance of file names, the linkage between files and the experimental programmes, and details of the processing would all be at risk. The ideal is that a future scientist would be able to reconstruct from the preservation archive as much as today's users know.

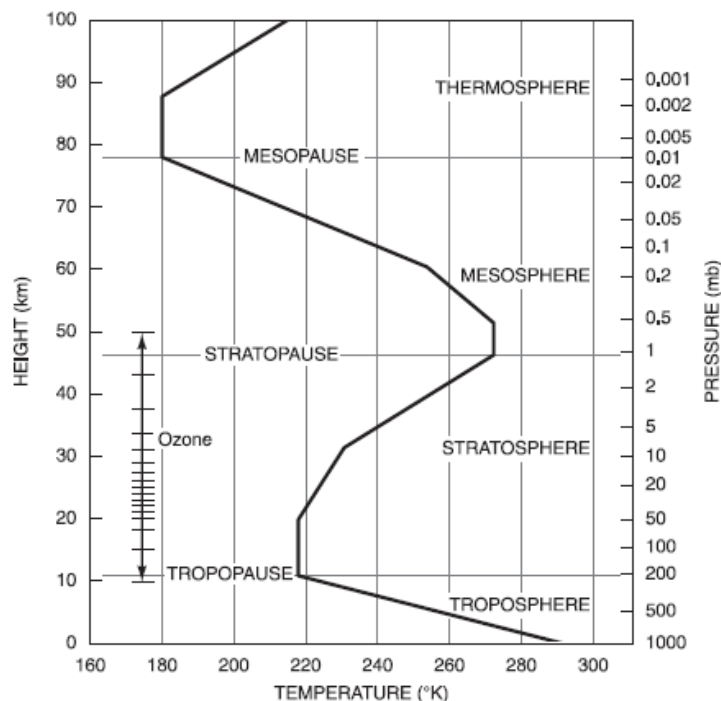


### 3.3 MESOSPHERE-STRATOSPHERE-TROPOSPHERE (MST) RADAR DATA (CCLRC)

#### 3.3.1 Introduction

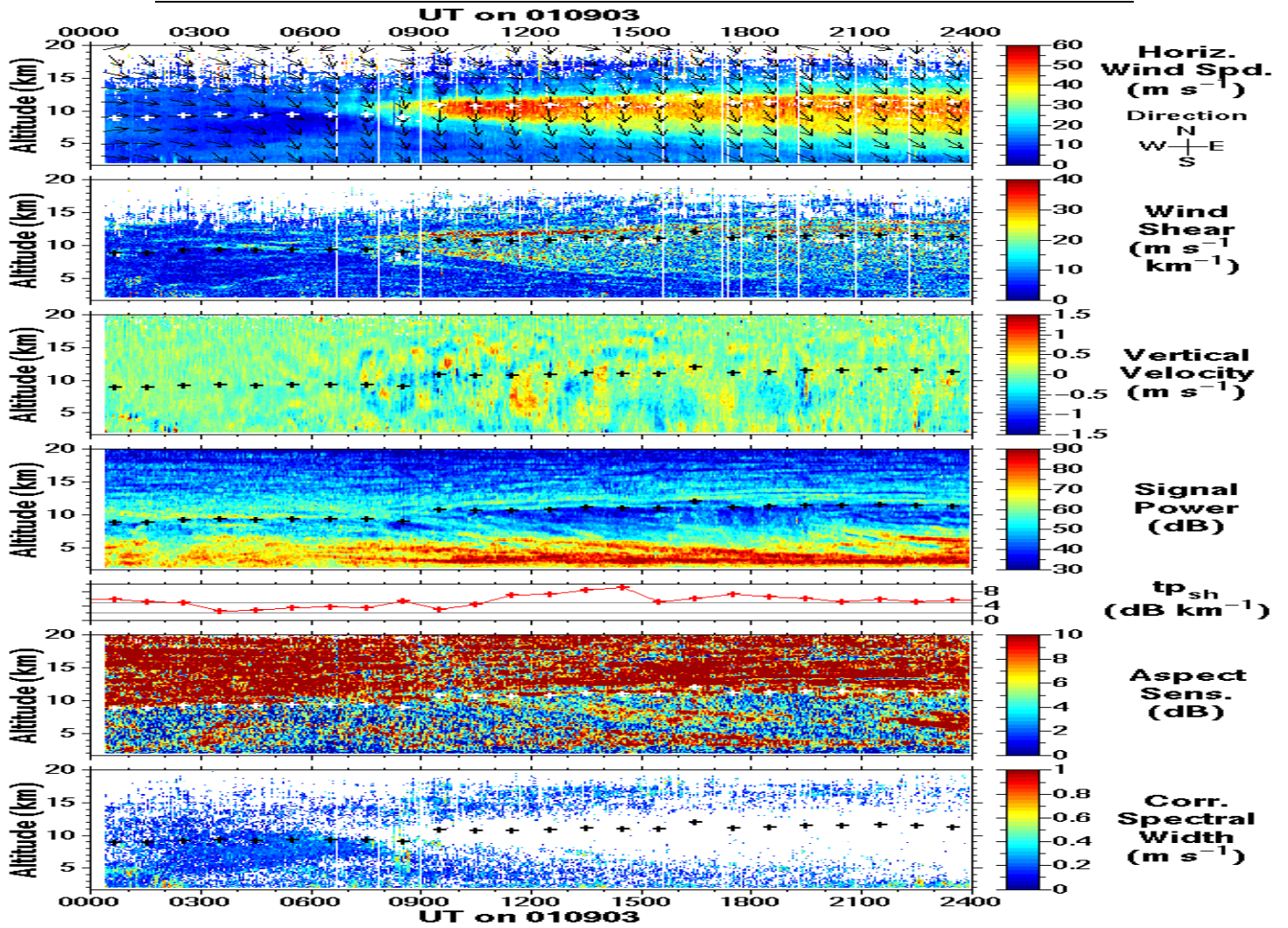


The MST Radar at Capel Dewi (near Aberystwyth, West Wales, and UK) is a 46.5 MHz pulsed Doppler radar ideally suited for studies of atmospheric winds, waves and turbulence. It is run predominantly in the ST mode (approximately 2–20 km altitude) for which MST radars are unique in their ability to give continuous measurements of the three dimensional wind vector at high resolution (typically 2–3 minutes in time and 300 m in altitude).



Doppler Beam Swinging (DBS) which involves making observations in a cyclic sequence of vertical and near-vertical beam pointing directions. The 'targets', from which small fractions of the pulsed radar signals are returned, are irregularities of atmospheric refractive index, which cause back-scattering (so-called 'clear-air' returns), and hydrometeors, which give rise to Rayleigh scattering. The scattered signal is Doppler-shifted according to the radial component of the target's velocity, i.e. that along the radar beams pointing direction. Profiling is achieved by sampling the radar return signals as a function of delay from the time of the transmitted pulse; the transmitted pulse length determines the range resolution.

Wind profiler radar returns are parameterised by their signal powers and spectral widths (i.e. the variance of scattered velocities about the mean) in addition to their Doppler shifts. This information can be used, under certain circumstances, to provide additional information about the atmospheric static stability (thus allowing monitoring of the altitude and sharpness of the tropopause), humidity fields and turbulence (of at least moderate intensity).





### 3.3.2 Preservation significance

#### Importance of MST dataset

- Contains data from the UK's most powerful and versatile wind-profiling instrument
- Provides information about atmospheric stability, turbulence, humidity fields, and precipitation
- Contains measurements of winds up to many kilometres from the ground (remote sensing)
- Contains a record of winds sampled continuously, with a cycle time of a few minutes over a long period of time
- Is a record not only of the horizontal but also the vertical air velocity which additionally has high temporal and spatial resolution





### 3.3.3 Preservation issues

CASPAR element	Current situation of testbed	<b>Preservation issues arising from general considerations and</b> <ul style="list-style-type: none"> <li>• <b>changes in hardware, software and Designated Community</b></li> </ul> <b>loss of sources of Information (including loss of host archive)</b>
Ingest	<p>Data file (IQ and spectral) regularly directly from MST station every day and deposited In appropriate directory. Further processing is performed to produce radial and Cartesian products. Scripts written in Python are used to deposit the files in appropriate directory and update appropriate databases.</p> <p>Plots of the Cartesian data are produced regularly by the project scientist and ingested into the archive.</p>	Need for incorporation of weblogs, technical specification, processing analysis and other technical provenance information into the archive and related to data in a meaningful way. Some digitisation may be required for info only available in physical form.
Preservation Description Information	<p><b>Provenance:</b> single instrument source, with version processing implicitly linked to file location. Descriptive information: web log (XML) of instrument operations/ performance.</p> <p>Much technical information on the instrument is in hard copy only and at preservation risk.</p> <p>Documentation on previous versions of signal processing including comparative analysis of the different versions by the project scientist and the Met Office.</p> <p><b>Fixity</b> reliant on Checksum which runs every 80 days</p>	<p>Provenance needs formalising and missing information needs to be captured.</p> <p>Ideally fixity would be assured from point of ingest.</p>





	<p>automatically if not manually performed. Back up copies can be retrieved from RSync disk back up and Atlas data store Data at risk of corruption during first 80 days in archive</p> <p><b>Reference:</b> file naming convention.</p> <p><b>Context:</b> BADC staff and scientist knowledge</p> <p>Finding Aids: Combination of Ingres database which holds data catalogue, Postgres which holds physical locations of data files. HTML and scripting.</p>	
Representation Information	<p>Format of file can be identified through a combination of file extension, physical location of file (as Indicated by Postgres database) and file name which indicates format. Textual description of the formats Is available on MST support pages</p> <p><b>Raw data</b></p> <p>IQ (In phase and quadrature) data - non standard binary format files format (textual description of format available)</p> <p>Spectral - Raw data which has undergone first stage processing to spectral non standard binary file format. (textual description of format available)</p> <p><b>Processed product</b></p> <p>Version 2 processed data product</p> <p>Radial data: Nasa Ames format (product specific textual description available)</p> <p>Cartesian data: Nasa Ames format (product specific textual description available)</p> <p>Version 1: processed data product</p> <p>Radial data: Nasa Ames format (product specific textual description available)</p>	<p>Formal language descriptions of the data formats and creation of data dictionaries for the parameters would allow independence from some of the access software.</p> <p>Preservation of processing version (version 0): a C program only working version currently compiled to run on Windows NT at high preservation risk as computer due to be retired Nov 06 and code impenetrable to even current project scientist</p> <p>Preservation of processing version 1 and 2 (currently requiring Matlab): either capture processing steps independently of Matlab or preserve Matlab (proprietary product)</p> <p>Preservation of processing version 3 (currently requiring Python) – either capture processing steps independently of Python or else preserve Python.</p>





	<p>Cartesian data: Nasa Ames format (product specific textual description available)</p> <p>Version 0: processed data product</p> <p>Time averaged radial data: non standard ASCII format (textual description available)</p> <p>Unaveraged radial data: non standard ASCII format (textual description available)</p> <p>Time averaged wind data: non standard ASCII format (textual description available)</p> <p>Unaveraged wind data: non standard ASCII format (textual description available)</p> <p>Time averaged power data: non standard ASCII format (textual description available)</p> <p><b>Quick Look plots:</b> graphic file png format generated from cartesian products using GNU plot</p> <p><b>Under development</b></p> <p>With version 3 binary file of cartesian and radial product will be produced in the NetCDF format in tandem with a Nasa Ames ASCII versions.</p> <p>Once Version 3 processing (Python producing NetCDF radial and Cartesian Product) is released a reprocessing of the archived Spectral file will be undertaken. In order to produce a consistent series of Cartesian and radial product. As they will be of higher quality and more easily supported.</p> <p>After a period of time the BADC would wish to dispose of the old higher level product but would like to maintain the ability to recreate it.</p>	
<p>Annotation</p>	<p>None</p>	<p>The ability to annotate the data with scientific theories on</p>





		atmospheric behaviour and to make the annotations searchable would be welcomed.
Packaging	Files stored in Atlas Petabyte Store – on Virtual Tapes	
Access	see finding aid under Ingest	Access and processing software currently require Python, INGRES and POSTGRES. Development of ontologies and data dictionaries from the CF standard names list is required.
Access control, including DRM	Access is restricted bona fide academic research or educational purposes and granted upon the approval of the project scientist after registration of interest. Access and quality control of data is subject to restriction imposed by NERC, BADC and the Met Office. See section 15 CCLRC_03 for further detail.	DRM/Access system which can deal with evolving access rights of data. Could something such as shibboleth be investigated to give access to data and associated copyrighted materials such as journals?
Higher-level knowledge	There is a great deal of associated information embedded in standard texts, journals, websites and specialist information of archive staff and supporting organisations. See section 2, 9 and 15 of [CCLRC-03] for further detail. The Archive manager is in the process of getting MST data parameters added to the CF standard names list.	Loss of information on web sites would affect usage of the data. Loss of expert staff would make use of data more difficult. Terminology has changed over time and this needs to be captured and formalised.
Virtualisation and representation information	Tabular data  Also some graphical digital objects produced by GNU Matlab, Excel, Notepad and IDL programs. GNU plot is method is currently used to produce the archive. Version 3 NetCDF files also require pycdf with unysis package are also required to access. The CDAT package of climate analysis is also commonly installed by end users access the netCDF binary files.	Tabular description of the raw and processed data would allow at least basic access. Preserving the ability to access both NSA Ames and NetCDF files.





Storage and storage virtualisation	Simple files, with automated backup systems to separate tape data store	
Preservation orchestration	None preservation is currently reliant on the state of the BADC	There are other MST communities in different countries with whom the BADC may wish to engage with
Authenticity	Relies on trust in BADC staff	





### 3.3.4 Preservation scenarios

#### 3.3.4.1 Changes in hardware and software

##### **Scenario 1: Change of Operating system so that GNU Plot (or other common plotting application) is no longer able to access NetCDF files**

A change of operating may mean that GNU plot will no longer run with (pycdf and unysis) or CDAT. The ability to preserve access to NetCDF files for common plotting applications is a requirement.

##### **Scenario 2: Change of operating system and loss of ability to carry out version 0 processing on Spectral data.**

A user requires the ability to create version 0 processed file to produce a consistent series of cartesian data. However with the retirement of the Windows NT computer (currently happening) the archive will no longer be able to produce the required file. The ability to do this should be preserved.

#### 3.3.4.2 Changes in environment

##### **Scenario 3: Changes in rights to data**

As the funding and policies surround institution such a NERC, The Met Office and the BADC evolve so do the rights of different users. Frequently, users have access to and use of different data sets and may also have entitlements to access copyrighted materials provided by commercial publishers. A Single Sign-On solution for the user would be ideal.

##### **Scenario 4: Changes in copyright ownership and legal restrictions on materials supporting knowledge base.**

A core requirement for the preservation of the knowledge extract from and usability of the data set are copyrighted or otherwise restricted materials which include:

- Journal Articles
- Bibliographies
- Books (standard texts)

Copyright restriction and ability to deal copyright issues is major reason for much of the material not being added to the archive.

The Archive would benefit from guidance as to which framework they should be operating under for the types of materials mentioned above and implications for cross border supply and International Property rights.

An alert service advising of key changes affecting such types of materials may be of use. As would automated alert to the clearance of copyright on specified (by archive manager) material such as key journal or texts so they be digitised and added to the archive or released from a dark archive.

##### **Scenario 5: Retirement of key personnel affecting knowledge base**

The texts relating contain the mature well established scientific theory. There is a very significant preservation risk with material this material as the majority of it is out of print with no record of numbers/holders of copies or organisational responsibility for preservation. The number of texts one would wish to store within the archive is dependent on the depth and breadth of knowledge one is seeking to preserve so it is difficult to provide and estimate of this.

Current end users have become dependent of subscription product such as web of science <http://scientific.thomson.com/products/wos/> to locate journal material. This constitutes a significant risk preservation risk ability to preserve such listings along with relevant subject indexing is something that requires further investigation.





The result is a need to create a bibliography in consultation with key people Atmospheric physics (see section 9 CCLRC\_03)

There many theories within the community which have not had sufficient research done to reach the journal literature. The annotation of these theories to the appropriate data would be welcomed as would identification of events

### **3.3.4.3 Changes in designated community**

#### **Scenario 6: Change in user community**

Use proposed methodology from section 2.2.3.3

As science evolves the reason why a scientist may wish to use this type of observational data will change. It is extremely difficult attempting to anticipate how a user community will change. However connecting this data into larger ontology which evolves over time for atmospheric science would assist the discovery and use of this data. The development of a data dictionary for parameters and an ontology which incorporates the CF standard names list.





## 3.4 EUROPEAN SPACE AGENCY EARTH OBSERVATION DATA TESTBED

### 3.4.1 Introduction

#### 3.4.1.1 Background

Science data in the Earth Observation (EO) field are generated by a great variety of passive and active instruments embarked on board of spacecrafts orbiting around the Earth. These data are used for scientific investigations and for operational and commercial applications. The preservation of such data, of all the ancillary data needed to process and/or convert them into useful information as well as of the processors and converters is vital for all projects/applications that need multi-temporal monitoring of the parameters influencing the investigations (e.g., progressive desertification in various areas, deforestation of Amazonian area, climate changes, coast erosion, urban expansion, Antarctica ice coverage variation, ozone hole, etc.).

ESA (European Space Agency) has acquired since 1979 data from a number of Third Party missions (US, Japan, France, etc.) and has launched ERS-1 (European Remote Sensing) in 1991 with 5 instruments, ERS-2 in 1995 with the same instruments as ERS-1 plus GOME (Global Ozone Monitoring Experiment) and Envisat in 2002 with 10 instruments. All the data coming from these missions are archived at selected facilities and accessible to the various user communities.

GOME is an ultraviolet and visible light spectrometer mounted on the ERS-2 platform in a configuration nadir-viewing (i.e. looking down vertically) at the Earth to derive a detailed picture of the atmosphere's content of ozone, nitrogen dioxide, water vapour, oxygen/oxygen dimer and bromine oxide and other trace gases.

The ERS-2 orbit ensures a global Earth coverage of GOME data every three days.

In the initial part of the mission data were acquired only by a set of ESA stations (Kiruna in Sweden, Maspalomas in the Gran Canaria island, Gatineau in Canada and Fucino in Italy). Today, following the failure of two redundant on-board recorders, data are transmitted in real time and acquired by a network of ground stations spread over the world.

GOME data, interleaved with data from other ERS-2 instruments, are recorded on board along the full orbit and then transmitted to ground in the so-called ERS-2 Low Bit Rate (LBR) downlink to ground stations.

Data de-commutation (removing the effects of the interleaving) at ground stations produces a data stream called EGOC (GOME extracted product files, also called Level 0 Data or Raw Data). The EGOC files are then transmitted in near-real time by the stations to the D-PAF (the German Processing and Archiving Facility) for further processing and generation of higher level products as Level 1 (radiances / reflectances) and Level 2 (trace gas amounts) data.

Note that EGOC can also be produced from level 0 tapes stored in the LBR archives and that at least one EGOC product file (covering 1 orbit) is required to generate one Level 1 Data product.

In the near future, all EGOC files will be converted from current format to SAFE format (Standard Archive Format for Europe, developed by the European Space Agency in the framework of its Earth Observation ground segment activities) and stored in a TBD ad hoc archive.

Summarizing, the end-to-end dataflow encompasses:

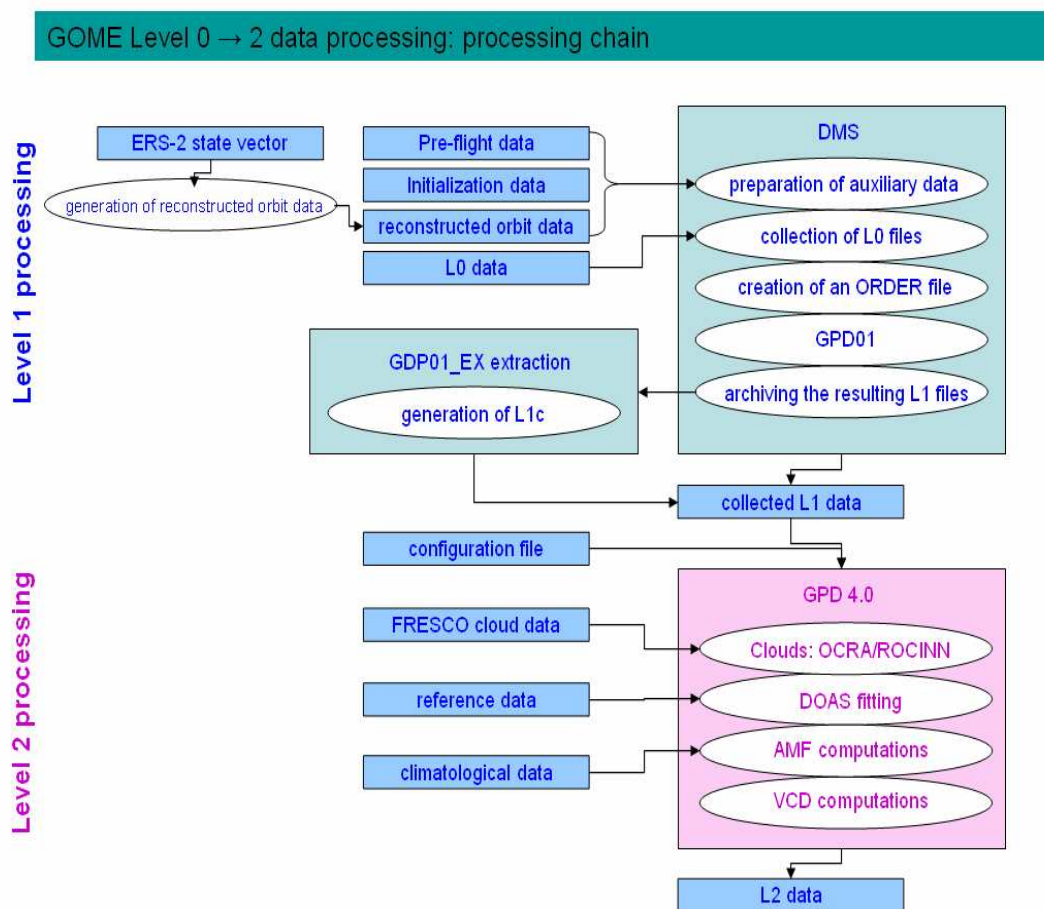
- sensing of the atmosphere with the GOME instrument and the generation of a corresponding GOME data stream inserted into the LBR data format;
- the (temporary) on-board storage of data;



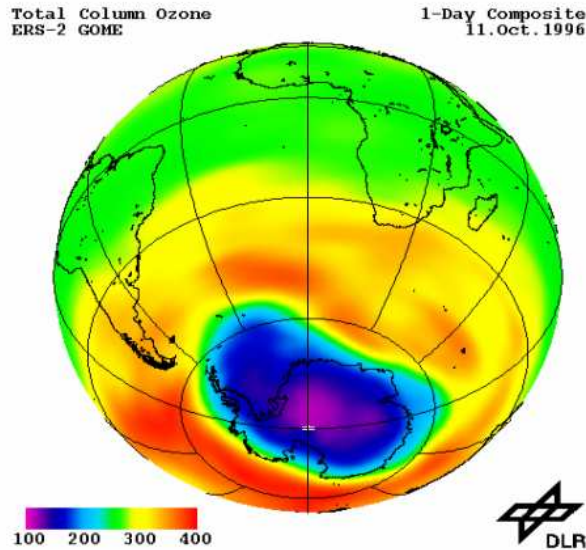
- their transfer to globally distributed receiving stations and next to various processing and/or archiving stations;
- the processing from level 0 (raw data), via level 1, to GOME level 2 data using auxiliary data and other data, information and knowledge (e.g., campaign surveys and documents) as needed;
- the access to GOME data of different levels by end-users via catalogue systems from corresponding archives.

This complex data flow has implications for preservation scenarios, and involves a large number of separate but interrelated data sets.

The GOME Level 0-2 processing chain is summarized in the following table to highlight the different types of data (satellite and ancillary) required for the derivation of useful information:



An example of a user product resulting from the above process is the Total Column Ozone here below showing the well known Antarctica ozone hole phenomenon:



With the above background in mind, it is clear that the preservation requirement regards both access and use of all satellite sensor data, metadata and ancillary data, knowledge and documentation, ontologies, algorithms and their evolution and validation history, various versions of the application software, etc., including access to other data sources (e.g. ground measurements used for calibration), models and whatever else used to support scientific research based on satellite data.

In order to assure a long-term data preservation, the data ingestion should be performed so as each object will be linked to all the objects needed to rebuild its complete history. Insertion of new information into the system needs to be easily correlated with the pre-existing context. In this way at the same time the system should create the preservation knowledge and perform the knowledge preservation.

### 3.4.2 Preservation significance

This science data preservation testbed will assess the Earth Science community requirements, develop the necessary specific services and prototype an Earth Observation science data preservation environment. In particular, the science data preservation testbed workpackage will carry out R&D activities specific to the testbed, will consider the infrastructure, set up the testbed and implement it.

The GOME dataset, because of its big total amount of information distributed with a high level of complexity, is a good case for prototyping in the CASPAR project.

Note that Earth scientists access not only the data from the satellite instrument to conduct their research. Often, they need access to complementary data from other instruments, including satellite sensors, air sensors, and ground sensors as well as to various models, algorithms, software and reports, publications and other documents that have information explaining how to produce useful research products, science results and so on. As such, the intended preservation infrastructure will deal with a mixture of selected simple/complex/on-demand objects based on GOME data, GOME data products, metadata available from catalogue databases (inventory and images), processing strings, calibration campaigns (with other instruments), datasets in other locations, large volumes of documentation regarding the instrument, the processing, algorithms, tools as well as with information about dedicated workshops, conferences and ESA-supported application projects, operational services deployed, and some special examples.



The key issue in these specific examples is the availability and reconstruction of detailed metadata associated with the very different types of objects to be managed: the testbed needs to ensure access to the above mentioned information, algorithms, tools etc..

The following table lists R&D activities that are considered relevant for the science data preservation testbed:

- The preservation life-cycle of scientific data: preparation of specific tools which will allow visualisation and navigation of the complex metadata associated with science data complex objects.
- Implementation of the OAIS Reference Model for Science data preservation.
- Compatibility with OGC efforts for geospatial data management.
- Evolution of the proposed SAFE (Standard Archive Format for Europe) archive format standards for Earth observation missions to include preservation requirements.
- Digital library integration.
- Metadata, Ontologies and semantics of science information objects: analysis of aggregation of satellite images and on-demand processed images from the Grid with corresponding text descriptions from news and from internal text archives as well as with other data available for the same period and same geographic area.
- The use of the Grid and on-demand (re)generation of information objects; some information objects are not stored as such but are generated on the fly. Preservation of such on-demand objects, that is, of the conditions under which they were generated and how to regenerate them is a specific research topic.
- Distributed partial copies of Earth science information objects; there are different problems related to availability of the 'same' data on different locations, e.g., authentication, authenticity and provenance. There is however also a problem related to partial copies. For example when satellite data are transferred from the satellite to the ground-station, it may happen that the reception is interrupted and received partially by different ground-stations. Research is needed to understand how to deal with these cases.

### Use cases

Typical use cases are (non-exhaustive list):

- access to/retrieval of level 0 satellite data (original and current format, etc.);
- access to/retrieval of various ancillary data (orbits, calibration, etc.);
- access to metadata (catalogue, browse, etc.);
- access to documentation and knowledge;
- access to methods and processors (source code, evolution, etc.);
- access to (e.g. via regeneration based on lower level products) higher level(s) of user products;
- access to and capability to rebuild and run a complete processing chain that allows for reprocessing of archived data to compare with recent capabilities and results;
- locate and extend the preservation data set with delta information (which were not archived in the first place) considered relevant for improved preservation;
- access to higher level products;





### 3.4.2.1 ESA test-bed

The testbed proposed by ESA covers:

- a) The definition, access, collection and/or identification of the location of a significant sample of the data set, documents, knowledge, algorithms/processors and methods representative of the GOME preservation scenario.

A possible choice of the preservation items is summarized in the following table:

#### GOME Level 0 → 2 data processing: preservation items

##### Preservation of Data

- Level 0 data (EGOC format)
- Level 0 data (SAFE format)
- ERS-2 State Vector → reconstituted orbit data
- Pre-flight data e.g. on ground calibration parameters
- Initialization data configuration parameters for GDP01
- Configuration file
- FRESCO cloud product
- Reference data (absorption cross sections)
- Climatological data

##### Preservation of Documents

- Product Specification Document of the GOME Data Processor
- GOME Data Processor Extraction Software User Manual
- GOME Data Product Improvement Validation
- GOME User Manual

##### Preservation of Processor

- GDP01 (GOME Level 0 → 1 data processor)
- DMS (Data Management System)
  - preparation of auxiliary data
  - collection of corresponding GOME downlinks (Level 0 data)
  - creation of an ORDER file
  - starting the GOME processor
  - archiving the resulting Level 1 products
- GDP 4.0 (actual L1 → 2 data processor)

##### Preservation of Methods

- ERS-2 state vector → generation of reconstituted orbit data
- GDP01\_EX extraction → generation of Level 1c (e.g. calibrated and geolocated radiances)
- EGOC → SAFE format converter

L0 → 1 data processing  
 L1 → 2 data processing  
 L0 → 2 data processing

- b) The integration in a GRID environment at ESRIN of the level 0 to level 1 DLR<sup>1</sup> software in order to exercise and demonstrate selected use cases such as:

- preserve/access the various data holdings and in particular a set of Level 0 data (simple objects);
- access Level 1 products (on demand objects), retrieving the Level 0 data in SAFE format and process them using a selected archived version of the DLR processing software;
- comparison of historical archived results with current capabilities.

Other proposals could be:

- to preserve “all needed” to perform format conversion (e.g. from the EGOC format to the SAFE format) when and if needed (and storage of products in a TBD ad hoc

<sup>1</sup> In the operational settings this software is running at DLR (Deutsches Zentrum für Luft- und Raumfahrt) in Germany.





archive);

- the development of the SAFE format even for the GOME level 1 data.

### 3.4.3 Preservation issues

ESA, the ERS-2 satellite owner, is the dataset producer via contracts with technical facilities managed by a group located in ESRIN, the ESA establishment in Italy.

At present processing chain, auxiliary data, processing algorithms, product inspection and quality control are provided by the facility contractor (e.g. DLR).

The GOME data set is unique; it provides more than 11 years global coverage.

GOME is complementary/precursor to EXOSAT GOMOS and to equivalent US instruments (TOMS).

The scientific community and principal investigators (PI) on a routine basis receive GOME data (e.g. KNMI and DLR) for their research projects.

Current preservation plans are based on the ERS programme declaration and ERS data policy which requires preservation of the data for a period longer than 10 years after acquisition. It must be pointed out that the current approach of satellite data owners (NASA, ESA, etc.) is to foresee an endless preservation of selected and/or strategic data sets; in other words nobody seems to be willing to purge data when preservation is technically feasible and economically affordable.

In early 2004 ESA has setup a project called HARM (Historical Archives Rationalization and Management), which aimed mainly at converting its historical datasets into a new modern format, based on the latest technologies and standards and able to ensure the long term preservation of its holdings.

SAFE (Standard Archive Format for Europe) has been designed to act as a standard format for archiving and conveying data within ESA Earth Observation archiving facilities and potentially with the cooperating agencies. SAFE intends to cope with the major constraints required for the packaging and the long-term preservation of Earth Observation data. A special attention has been paid to fit the Open Archive Information System (OAIS) reference model and related standard of the CCSDS such as the emerging XFDU packaging format.





CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and <ul style="list-style-type: none"> <li>• changes in hardware, software and Designated Community</li> <li>• loss of sources of Information (including loss of host archive)</li> </ul>
<i>Ingest</i>	<p>GOME Level 0 data as archived EGOE files or to be extracted from the LBR format.</p> <p>Ancillary, auxiliary data and metadata in various formats and locations.</p> <p>EGOE files will be converted from current format to SAFE format.</p>	<p>Reformatting of data will have as a consequence the re-transcription of all data. This may imply also change of technology and/or location of archive, including file names and directory structures.</p> <p>Also documentation will have to be updated accordingly.</p>
<i>Preservation Information</i>	<p><i>Description</i></p> <p><b>Provenance:</b> source implicitly inserted in data files.</p> <p><b>Fixity:</b> relies on stability and correctness of systems. Monitoring degradation or alteration.</p> <p><b>Reference:</b> file naming convention.</p> <p><b>Context:</b> metadata collect all info about the dataset.</p> <p><b>Finding Aids:</b> placed in the database used by catalogue and inventory.</p>	<p><b>Provenance:</b> There is a substantial variety of sources due to the huge number of information types to be preserved. Provenance should be formalised including source of data and method of transfer. For processed data at all levels the processing steps and algorithms versions should be detailed and preserved with the associated workflow.</p> <p>A SAFE product gathers product basic information (the EO data is completed by Fixity Information and a fully Representation Information); SAFE metadata tags collect the historical information dedicated to the maintenance/traceability of the product and contain information about the quality of an EO dataset.</p>

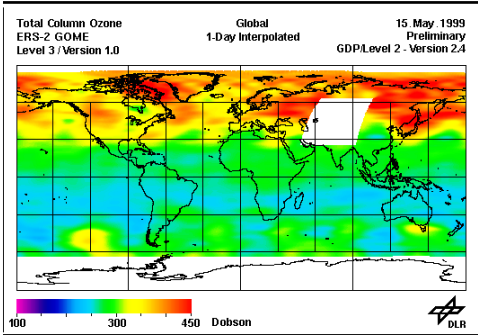




<i>Representation Information</i>	Information is provided with the documentation as listed in <a href="http://earth.esa.int/services/esa_doc/doc_gom.html#">http://earth.esa.int/services/esa_doc/doc_gom.html#</a> and <a href="http://earth.esa.int/SAFE">http://earth.esa.int/SAFE</a> . All data set are characterised by their format description and annotations.	There is room for standardising representation. SAFE format is a step in this direction. Effort should be made also to cover metadata and ancillary data. Many of the files are described using a proprietary system from GAEL. This data description technique needs to be adequately documented to remove dependence on proprietary software. SAFE is a derivative of the XFDU work which is in the process of being standardised.
<i>Annotation (could be regarded as a special type of Rep. Info.)</i>	None	
<i>Packaging</i>	Many files in simple directory structures. SAFE is a packaging technique.	
<i>Access</i>	Metadata are accessible via WEB based finding aids. Level 0 data and ancillary data are accessible to ERS project internal use via computer transfer protocols and to restricted number of external users on DLT or CD. Level 1 data are available via ftp, on CD and DLT media (covering the time period from August 1995 until today). Level 2 data products are available via ftp server and also on CD-R containing one month of data.  In addition, Level 2 data products are available on the Level 1 CD-R, along with the Level 1 data products (if already processed) and associated software and documentation.  Documentation is available through ESA and DLR portal.	Data access is ruled by the ESA ERS Data policy. Also implementation and procedures are subject to ERS project resources.
<i>Access control, including DRM</i>	In principle there is no restriction access to all preserved information. In practice, since raw data handling requires project specific tools and information/skill as well as expensive equipment, level 0 data are only accessible by Principle	Data policy is subject to changes by decision of ESA member states.





	Investigators and ESA internal users. No access restriction on fast delivery products, level 1, level-2 or higher level products.	
<i>Higher-level knowledge</i>	Information is provided with the documentation as listed in <a href="http://earth.esa.int/services/esa_doc/doc_gom.html#">http://earth.esa.int/services/esa_doc/doc_gom.html#</a> and <a href="http://earth.esa.int/SAFE">http://earth.esa.int/SAFE</a> .  No specific support is provided during data usage by scientists.  Only issues of data quality, instrument status and algorithm updates are covered.	Documentation is in evolution and updating as needed.
<i>Virtualisation and representation information</i>	As an example we provide a level 2 GOME Satellite instrument data itself:   <p>This is a mosaic composed by the data retrieved over 14 orbits per day with interpolation of data values to fill the gap between two quasi-adjacent geographical areas covered by the satellite ground track. On that date the white zone over Asia indicates absence of useful data.</p>	Products are in evolution both in terms of quality and of type.
<i>Storage and virtualisation storage</i>	Storage of satellite raw data is performed at dedicated centres in form of digital tapes in a robotized environment. Other data/metadata are stored as simple files.  Some rationalisation could be performed at the end of the ERS mission to concentrate all data set in a dedicated single facility.	Need to support mass migration between large scale systems such as the robotic environment mentioned. This could involve capturing internal details (catalogues, metadata) of the robotic system.  Preservation policies of such large scale systems





	A GRID environment and/or the use of a digital library (as seen in the FEDORA or DILIGENT ESA projects) may be used.	should also be captured. Preservation implications of import into FEDORA or DILIGENT should be considered.
<i>Preservation orchestration</i>	Due to redundant storage in different locations (though in different form and support) there is no risk of data loss.	Change could be introduced as result of recommendations from CASPAR project.
<i>Authenticity</i>	It is based on the reliability of the preservation environment (specialised centres) and in the rigorous procedures during re-transcription in case of change in technology or of changes in formats.	Current systems are fully documented and put under configuration control. Any future change in hardware, software, algorithms and procedure should undergo strict procedures of configuration control/documentation. Effect of changes will be monitored by comparison of reference datasets or products before and after changes.





### 3.4.4 Preservation scenarios

In the context of the ESA test-bed the scenario envisaged includes the conversion of GOME EGO files to SAFE format and the reprocessing of some data up to level 1. This will imply the re-transcription of the raw data starting from either the EGO files stored in DLR archive in Germany or from the LBR data streams archived in DLT in Matera (Italy).

The activities to be performed are:

- Software development for the reformatting of the GOME raw data in SAFE format and full dataset re-transcription;
- Installation of DLR processing software in a GRID environment at ESRIN establishment in Frascati;
- Processing of a selected portion of GOME data from level 0 to level 1;
- Definition and implementation of a strategy for the preservation of data and to comply with the objectives defined in the CASPAR context.

The choice of the GRID environment as basis for the middleware seems to be good because of its independence from operating systems and its catalogue functionalities. Moreover, a digital library should have the optimal characteristics to perform the content management: the EU project DILIGENT should have some features that could be useful for the CASPAR project purposes and they will be carefully examined.

The activities to be performed also envisage measures to select the status of storage media, to achieve hardware modernisation and implement software adaptations; the compilation of adequate project documentation and the adaptation of user catalogues and documentation for easy data access will also be pursued.

#### 3.4.4.1 Changes in hardware and software

Change of

- hardware
- operating system
- compilers/libraries/drivers

affecting ability to run

- the Data Management System
- the GOME Data Processors
- the format/auxiliary data converters.

Need to preserve the ability to generate on demand Level 1 or 2 data starting from Level 0 data. Either software and the ability to run it must be preserved or sufficient documentation to allow for the development of a clone system.

#### 3.4.4.2 Changes in environment

Change of software copyright affecting ability to read SAFE data.

Need to preserve the ability to access and use SAFE data. Sufficient documentation to allow for the development of such a software must be preserved.

#### 3.4.4.3 Changes in designated community

Change in user community

Use proposed methodology from section 2.2.3.3





## 4 THE ARTS DOMAIN

### 4.1 THE IRCAM TESTBED

#### 4.1.1 Introduction

Performing arts including electronics concern all arts in the field of performance requiring electronic device of any kind, either analogue or digital, used for signal processing or symbolic calculation. Performing arts include of course music, but also dance, theatre, video, interactive installations, etc. They may require human performers or not. Here are a few examples (not exhaustive):

<i>Configuration</i>	<i>Example of work</i>
Tape diffusion	<i>Gesang der Jünglinge</i> by Karlheinz Stockhausen, opus 8, for electronic and concrete sounds (1955-1956)
Solo instrument and live electronics	<i>Anthèmes II</i> , by Pierre Boulez, for violin and live electronics (1997)
Soloists, ensemble and live electronics	<i>Répons</i> , by Pierre Boulez, for six soloists, chamber ensemble and live electronics (1981-1984)
Dance Performance	<i>L'écarlate</i> , dance performance designed by Myriam Gourfink, choreographer, music by Kasper Toeplitz (2001)
Theatre with Electronics	<i>La traversée de la nuit</i> , by Geneviève de Gaulle-Anthonioz, stage production using realtime neural networks and multi-agent systems (Christine Zeppenfeld, Alain Bonardi, 2003).
Musical and video performance	<i>Sensors Sonic Sights (S.S.S.)</i> , music/gestures/images with Atau Tanaka, Laurent Dailleau and Cécile Babiolo (performed since 2004)
Installation	<i>Elle et la voix</i> , virtual reality installation by Catherine Ikam and Louis-François Fléri, music by Pierre Charvet (2000)

#### Performance materials and logic

Performances including electronics are based on **various materials**. These materials belong to three types :

1. **A priori texts**: for instance, librettos, text fragments, scores, stage indications, etc. They may be prescriptive (for instance: sing with forte dynamics) or descriptive (move from the right to the left of the stage). They are the primary sources of the performers' work.
2. **Active modules**: either hardware (analogue devices) or software (patches), they provide control and signal processing functions.
3. **Technical documentation**: schemes, instructions (for instance, for the audio diffusion), drawings, etc.

The Ircam testbed will be concerned with mainly solo instrument and live electronics, with or without ensemble, but can also include Dance Performance or Musical and video performance. The Ircam archive consists of more than 450 works produced since its





creation in 1974. From these 450 works, more or less than 70 works can be considered as very significant, and are the object of new performances, specific studies, or even for some of these serves as references for young composers.

#### 4.1.2 Preservation significance

Electronics for performance are designed according to two paradigms : signal processing and control interfaces.

##### 4.1.2.1.1 Signal processing

One of the key points of electronic music is the deep influence of signal processing. In the past, composers would use analogical device and connect them together, having the output of one of them linked to the input of another one. Today this is the same apart from the fact that nearly all artists use graphical languages for signal processing that are based on the same paradigm: boxes that provide signal transformations (in a broad meaning) are connected together. Examples : Max/MSP/Jitter, PureData, Isadora, etc.

##### 4.1.2.1.2 Control interfaces

Control interfaces are based on one main element: devices from the industry (microphones, MIDI keyboards...), very frequently complemented by specific processing implementations generally using the same processing engine as noted above (including Max/MSP/Jitter, PureData, Isadora, etc.). Sometimes specific control devices are developed for specific purposes. An annual event (New Interfaces for Musica Expression) is devoted to this activity.

Concerning the representation of control – especially Midi Control - in graphical languages, it is in a way inspired by such methods as Petri networks and Graphcet (developed during the 1960's). More recently the field of man to computer interfaces has influenced further development in all graphical languages (Max/MSP/Jitter, Pure Data, etc.), including new objects for interface purposes (multi-parameter graphical interfaces, etc.)

#### Performance works including electronics are fragile

They are very sensitive to:

- *the wearing out of storage support media* :
  - CD-ROM, USB keys, hard disks, DAT tapes, etc., are fragile storage.
- *the obsolescence of technical standard* :
  - Use of end-user configuration patches using proprietary software and/or hardware technologies, and use of binary and especially proprietary file formats are prone to obsolescence. Formats like ADAT for multitrack tape music are almost obsolete.
- *the lack of care in the preservation of the electronic part*:
  - For many years, libraries would keep scores, but composers were responsible for the preservation of their own electroacoustic or live electronics parts. See the example of Donemus/Near. It is often the case that documentation about live electronics is insufficient.
- *the practical conditions of performance*
  - Changing performing conditions may produce unexpected results. For instance performing with slightly different device or in different places always needs time to control as much as possible and often requires modifications.
- *when applicable, the empirical nature of making patches.*





- The complexity of patches sometimes makes them very unstable. For instance, in Max/MSP software, it is well known that priority processing depends on the position of objects.

### **Performance works including electronics look for sustainability**

The worse situation for all these artistic works is the impossibility of re-performance (for various reasons) after the creation. Therefore, institutions concerned and composers are interested in sustainability. It paradoxically means preserving authenticity and at the same time enabling possibilities of evolution.

#### *The preservation of a level of authenticity*

Each time a work is played again, performers face the issue of authenticity, trying to set a kind of loyalty, in reference to the materials and to previous reference performances (maybe belonging to opposite interpretation traditions), famous or not.

#### *Possibilities of evolution*

Composers often want to modify their works, or performers intend to adapt a work to other configurations. Maintenance of software patches is a very difficult task, since they are not structured as programs for instance; slight modifications of a patch a few months after completing it may become quite laborious.





### 4.1.3 Preservation issues

CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and
Ingest	<p>Three stages :</p> <p>1) elements from the production are gathered by the person in charge of the production (the "musical assistant"). More information can be produced at this stage : diagrams showing details of installation, information regarding Intellectual Property, screen copies showing details of installation, description of installation process...</p> <p>2) whole information is packed as an archive (.zip) and uploaded on the server using an online tool (<a href="http://mustica.ircam.fr">http://mustica.ircam.fr</a>)</p> <p>3) editorial metadata (composer, name of the work, version number...), are added to the database in order to make possible queries.</p>	<p>[PAIMAS] issues</p> <ul style="list-style-type: none"> <li>- Packaging of all information related to a version of a work in one stand alone archive</li> <li>- Metadata should be packaged in the archive</li> </ul>
Preservation Description Information	UNKNOWN	<p>Provenance</p> <p>Fixity</p> <p>Reference</p> <p>Context</p>





Representation Information	Formats determined by file extensions Semantics is implicit	Proprietary formats are an issue. Proprietary software, with time limited licences is an issue. Behaviour of digital content, plus required hardware and software, is an important aspect which needs to be captured.
Annotation	None	
Packaging	Details missing (e.g. directory structure and expected contents) for ISO 9660 image which is archived	AIP needs to be defined
Description Information		
Access	Access to archive and ability to download is ensured by an online web interface ( <a href="http://mustica.ircam.fr">http://mustica.ircam.fr</a> ). Ability to locate the works by : - Author - Name of the work - Version number Ability to download the entire work as an archive or a part of the work	
Access control, including DRM	Access control is based on a simple login/password No DRM is implemented at present time.	Issue about longevity of usernames/passwords – how long are they valid. What happens after the death of the authorised person?
Higher-level knowledge	Today no high level knowledge is extracted from the production process in a systematic way.	Issue about how to describe the intentions of the composer regarding time processing, timbre result, spatialization, independently of the actual implementation of the digital process.
Virtualisation and representation information	Potential virtualisation: audio streams, digital processing algorithms	Issue: how far can the performance be regarded as a “workflow”?





Storage and storage virtualisation	No specific storage is used at present time.	
Preservation orchestration	None	
Authenticity	<ul style="list-style-type: none"> <li>- Examples of results (sound samples...) are stored with the archive in order to know if the result gained in the performance is correct towards the intentions of the composer</li> <li>- Evaluation of authenticity of reperformance is not trivial, and should be based on critical studies mad by authorities (musicologists...) and should take into account evolution of the user's perception with time, as well as the evolution of the production staff (including performers).</li> </ul>	<p>Issue about authenticity of performance: could involve comparison with</p> <ul style="list-style-type: none"> <li>- Stored audio excerpts of previous performances</li> <li>- Description of the intentions of the composer</li> <li>- Continuous feedback on successive performances of the same work</li> <li>- Tools for helping to evaluate the authenticity of a re-performance (comparison tool for example)</li> </ul>





#### 4.1.4 Preservation scenarios

In the following scenarios, we make distinctions between different elements of the digital part of a work.

These parts are:

- The equipments : input devices (microphones...), output devices (synthesizers...)
- The hardware running the operating system (e.g. a Macintosh)
- The system software where the real time application is running (e.g. Mac OSX)
- The real time processing engine where the patch is running (e.g. MAX/MSP, PureData, EyesWeb...)
- The "patch" itself where the logic of the work is implemented

The performers and the production staff in charge of the new performance are also to be taken into account in these scenarios.

##### 4.1.4.1 Changes in hardware and software

###### Scenario 1: Changes in the equipments – industrial devices produced by industry

These changes could affect input devices such as microphones or output devices such as synthesizers

The questions are :

- are the characteristics of the original device sufficiently known?
- is it possible to find an equivalent device?
- can the CASPAR process help to identify such a device?

If the result is negative, this scenario can be assimilated to the second one below.

The CASPAR process can help solving these issues by, for example, investigating the development of a model of description for such equipment, or adopt a model of description provided by the industry. CASPAR could also develop a mechanism for adequately identifying devices according to the required characteristics. A mechanism for helping to identify changes in perceived results, and preserving authenticity is also needed.

###### Scenario 2: Specific equipment no longer available

The change can affect mainly devices (control interfaces) specifically developed for specific purpose - like MIDI flute.

The questions are :

- are the characteristic of the device sufficiently known?
- is it possible to develop an equivalent device?
- can the CASPAR process help to develop such a device?

The same considerations apply as for Scenario 1 above.

###### Scenario 3: 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 :

- does the changes affect the behaviour of the patch?
- does the CASPAR process help identify which behaviour are affected by the changes?
- 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: changes in the availability of the real-time processing engine.**

The changes can affect the the real-time software (e.g. MAX/MSP) 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.

**4.1.4.2 Changes in environment**

**Scenario 5: changes in the behaviour of performers.**

The changes can affect the interactions between the performers and the "patch" (the logic of the work).

The questions are :

- does the CASPAR process help identify which behaviour of the patch is affected by the changes?
- does the CASPAR process help identify if these changes affect the authenticity of the work?

The CASPAR process can help solving these issues by investigating the development of a model of description for the interactions between the performer and the logic of the work (the patch), and develop a mechanism for identifying affected behaviours. A mechanism for helping to evaluate the authenticity of the reperformance is also needed.

**Scenario 6: changes in the production staff (e.g. directives).**

The changes can affect the perceived results of the performance.

The questions are :

- does the CASPAR process help identify which perceived results of the work are affected by the changes?
- does the CASPAR process help identify if these changes affect the authenticity of the work?

The CASPAR process can help solving these issues by developing a model of description for the perceived results of the work, and develop a mechanism for helping to evaluate the authenticity of the reperformance.

**Scenario 7: changes in the legal environment (e.g. patents on digital process).**

The changes can affect the rights to migrate, emulate or virtualize a digital object (due to reverse engineering).

The questions are :

- does the CASPAR process help identify a similar digital object which can be used instead of the original one ?

The CASPAR process can help solving these issues by developing a classification of digital processes, a kind of "organology" of audio processes.

**4.1.4.3 Changes in designated community**

**Change in user community**

Use proposed methodology from section 2.2.3.3





## 4.2 THE INA TESTBED

### 4.2.1 Introduction

The INA GRM testbed concerns Acousmatic Music, which can be defined as a musical work done through technological manipulation of sounds and existing as a fixed media. The preservation of Acousmatic Music works is not the only issue, it concerns preserving the production environment and their associated elements as well the performance conditions in which the work is proposed to an audience.

Increasing access to the whole environment of musical works is demanded by the public as well as by specialists; through an evolution of practices in which music creation is today a feasible issue for many people—even without specific training, willing to understand the creation processes and the evolution of ideas.

#### 4.2.1.1 The Acousmatic Music production scenario

Acousmatic Music<sup>2</sup> is a musical domain that started existing at the end of forties initially in Europe but then in many other countries and regions of the world. It is a consequence of the increasing use of technology in music, and finds its origins in the electric instruments and devices invented during the first half of the Twentieth century, as well as in cinema and radio sound experimentation that these two media developed in a continuous quest for new sounds and environments.

Its main concern is the creation of musical works composed for a fixed media; in which recorded sound is manipulated, modified and recombined in order to obtain musical structures using any kind of existing sound however based in its beginnings in the musical structures of the past. The result is a recorded media, which contains the musical work as the composer imagined it in all its details; it is thus a media dependent musical work, which has followed through its evolution the technological evolution of the compositional environment. In its origins, the used media was black records (Shellac), then magnetic tapes in different standards and formats, and in the last fifteen years digital media (DAT, CD, DVD, SDLT). Its main characteristic is that it is a self-describing artistic work, the media contains the work of art with all its necessary parameters and information, it only needs to be accessed through an intermediating system called a player (tape-recorder, CD player, etc.) that transforms the coded information into audible musical sensations. The problems related to the preservation of Acousmatic works is strongly related to Audiovisual preservation, but with particular aspects concerning sound quality and the way in which the music is presented to the listener during the concerts.

From a perception point of view, the main characteristic concerning Acousmatic music is the fact that the listener has no or little knowledge about how the sounds were produced and what do they eventually represent. The perception of Acousmatic Music relies in the previous listening experience of the listener and the capacity all individuals have of reconstructing the origin of whatever sound is heard, even if it is an unknown source. Thus, listening Acousmatic Music becomes a double experience of image reconstruction of possible sources and production situations as well as the musical organisation of those perceived structures. Any sound can be used in this domain, which is open to many experimentations and different kind of performance situations: pure Acousmatic Music for Concert, accompanied by performing instruments, with associated images, etc.

---

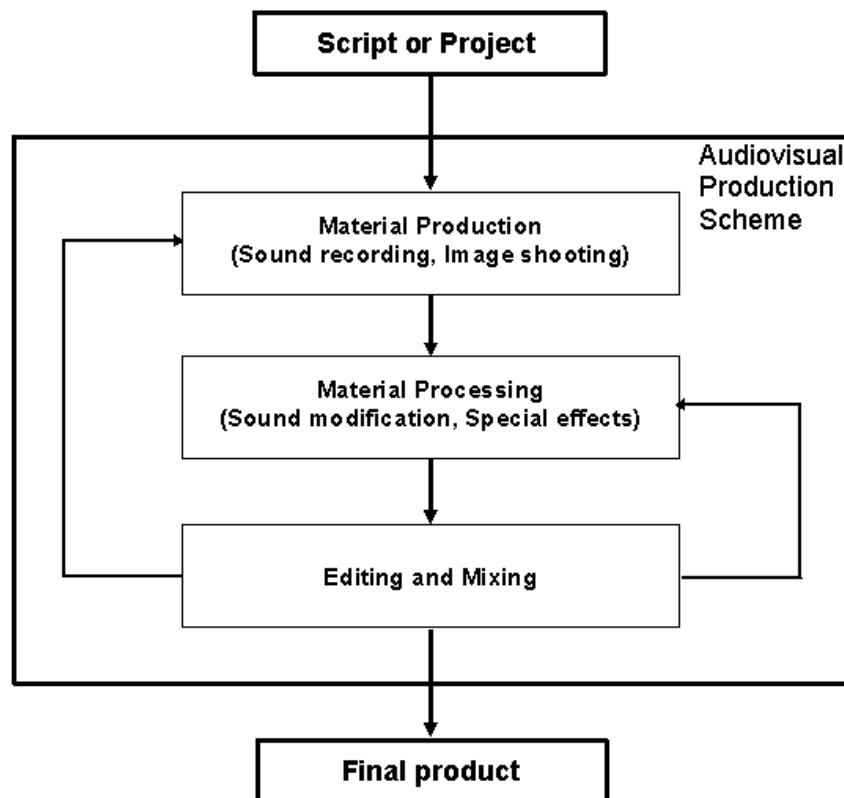
<sup>2</sup> The term « Acousmatic » means: what can be listened without seeing the causes. The term seems to have been invented by Pythagoras, who used to keep beginner students behind a curtain during the first two years of their apprenticeship so they would concentrate on the conveyed information and not on the mimics of the teacher. The term was reintroduced in 1955 by Pierre Schaeffer to describe the particular musical situation in which musical sounds could be heard without perceiving any visual cause. The term is explicated clearly by François Bayle in: Bayle, François, *Musique acousmatique, propositions... positions*, Éditions Buchet Chastel, Paris, Paris, 1993





The origins of Acousmatic music are to be found in cinema and audiovisual production. Cinema, Radio and finally Television, due to the complex production teams and technology that were needed in order to create a work, have used since the beginning a well defined scenario, that ensures a correct collaboration among the different intervening teams and organises the process through time.

The general production scenario used by audiovisual can be described through the following structure, which describes the actions through time:



Audiovisual production has thus structured the activities in order to optimize the presence of individuals and technology at the same moment in a same place. The scheme also responds to a logical sequence, in which initially all the necessary material is produced, and then selected items are modified and finally all the elements are combined together to produce the final work. There may be one or more feedback loops in which new material and new processing is needed in order to complete missing elements.

**The Production Scheme applied to Acousmatic Music:** In each audiovisual production chain, the importance of each action will be different, depending on the technical complexity of the task and the needed human interventions. Acousmatic production originated in the Radio studios, which were the only places where the necessary technology and know-how could be found. Initially music was done with two human components: the composer and the technician who would manipulate the machines. Progressively the composer became competent technologically and at the end of the sixties he started being the only operator in the process. With the arrival of digital technology, and the necessary skills needed by modern composers, the whole process can be done in home studios.

However, the great lines of the scheme have not changed, with a very strong stress put in the second action of the scheme: Sound processing. The Audiovisual production model, when applied to Acousmatic Music, can be presented under the following actions:





1. **Material Production:** Material can be produced either by sound recording either through sound synthesis. This action determines what will be the essential sound environment of the work. Generally a set of few coherent sequences are recorded or generated that will serve as master material for the rest of the work. Any sound can be used to generate the work, it is however an essential aspect in the composers project, it consists not only in finding an object, instrument or situation that will produce an interesting Sound, it implies that the person producing the sound will do it in a specific way, sometimes following a script, with many essays and errors. When Acousmatic sounds are mixed with performers, generally the composer uses sounds produced by the instrument so to create a strong coherence between the two worlds. Usually composers may use two or three different sets of sounds, of different nature and with a different spectrum and morphology (the range of possible sounds is too large to be mentioned, it may imply huge sounds from nature to infinite micro sounds). Technically this is a very delicate process in which sound has to be recorded in the best technical conditions so to ensure that the recorded elements convey only the needed material and not unwanted parasites. Once this stage finished, material is listened thoroughly, organised in sequences and described.
2. **Sound Processing:** This is a very important action during the process and, in comparison with other media, one of the most extensive. Most of the material of the work will be generated during this process using different Sound Processing Tools, which will apply different transformation processes to sound. The main transformations can be catalogued in two categories:
  - a. Modification of the Time organisation of the sound phenomena (delays, multiplication, speed variation, reversing, micro-editing, shuffling, time-stretching, etc.)
  - b. Modification of the spectral distribution or organisation (analysis and re-synthesis, filtering, transposition, hybridising, etc.)

The Sound Processing action generates large quantities of material derived from the initial material produced in the first phase of the process. It guarantees a “genetic” relation between the original recordings and all the other sounds. Normally tenths of different sequences can be obtained from an original recorded sound or sequence. Much more material than needed is produced so to dispose of a very large array of sequences among which the composer may choose the elements for his music. The material used during this stage to generate the material can be extremely diverse and hybrid, ranging from old analogue material so to obtain time-dated sounds, to the most modern software and processing machines.

3. **Editing and Mixing:** The final assembly phase of the work, in which all elements are put together following the composer’s project. All the previously produced material is listened thoroughly and annotated so to obtain a documented description of it. The composer starts assembling the material in a specific environment called a Sound Sequencer (many exist on the market), which permits the user to edit the sounds (select some specific regions within sequences), modify the level of sounds so to ensure a proper setting among them, and superpose them so to build complex sounds or to assure continuity. It is a slow and handcraft like process in which little by little the pieces of a very complex puzzle are put together. Composers may work on an already defined project or script, which they follow in detail, or either work reacting with what they listen and adjusting the project in function of the sounds and the relations that may surge among them. It is important to remark that, unlike instrumental music in which the sound material (the instruments) is constant and can be dealt abstractly and thus premeditated; when dealing with completely new sounds, abstraction is more complex and unexpected relations surge among sounds when associating them to convey a musical idea.
4. **Feedback loops:** The actions may be reintroduced during any of the stages:





- a. If during the Sound processing stage the initial material is considered as not being efficient enough the recording process or sound generation process may start again
- b. It may happen during the Editing and Mixing stage that not enough material is available, new processing may be needed and eventually new recordings or processing
- c. Sound processing may be added during the Mixing stage, without generating new material but through local processing on already mounted sounds to enhance some aspect of it
- d. Finally, all the stages may be condensed when a composer wants to do the recording, the processing and the mixing at the same time, hoping frequently between one or other action

**Conclusion:** The described actions, integrate the observed logical process for composing Acousmatic Music. The main reason underlying the fact that they are a succession of actions is that each of them implies a specific behaviour and concentration. For the **Recording** action, stress will be put on the quality of the generated sound and the efficiency of the object or instrument manipulation that will produce the Sound sequences. During the **Processing** action, stress will be put in the parameters that will define the modification of Sounds; different sequences will be generated with different sound processing techniques. Finally during the **Mixing** action, the main objective is the global result. During the previous stages the composer deals with potential musical sounds, during this final process the sounds become a part of the music and have to be carefully calibrated.

**Concert presentation scenarios:** Different use scenarios are possible once the music composed; it will ultimately depend on the initial project (concert, multimedia, installation). Generally Acousmatic Music is composed for the concert, which means it will be presented in a large space with an audience and a loudspeaker environment. The previously described composition scenario has not changed much through time; the actions are regularly confirmed by the procedures composers apply during their composition process (procedures regularly checked with composers working in the studios of GRM). What has mainly changed through time is technology, the media on which the works are recorded and the concert situation in which the work is presented to a public.

The final media used in Acousmatic Music has followed the evolution of domestic listening formats for recorded music, with some deviations in the last ten years:

1. **Monophonic** recordings, were the only possible way of listening until the middle of the fifties
2. **Stereophonic** recordings exist since the fifties and are still largely used, they will probably continue for a long time since they mimic our human bi-aural listening system
3. **Multiphonic** recordings, starting in the sixties but only becoming widely used in the nineties, when an 8 track standard became very popular as a circular proposition for listening music. A domestic version of this surrounding situation was brought in by the Home cinema 5.1 systems (derived from cinema uses), which, with a reduced amount of loudspeakers, suggest a convincing surrounding listening. Many composers still work with the 8 tracks standard, very practical for organising the movements of sound in space (8 track listening is not a domestic array).

The recorded media has then to be adapted to a concert situation, which may range from a small concert hall with less than a hundred people to several hundreds or a thousand. Loud-speaker orchestras have been developed since the sixties, sometimes receiving the name of "Acousmoniums", these are complex loud-speaker installations, sometimes ranging up to one hundred loud-speakers distributed all around the hall, which permit a complete control of the behaviour of the sound in the concert hall. Most of these systems are hand-controlled, and specific dispositions are done in function of the hall and the music characteristics.

#### 4.2.2 Preservation issues

It may seem that the essential issue for an Acousmatic work is to be preserved as such; in the same way a traditional instrumental score seems to convey the necessary information for music performance





(which it does not, it conveys partial information, the missing information is reconstructed through use tradition). However, the situation is much more complex due to the characteristics of the production environment, which interfere with the nature of the result and the number of possible elements to be preserved.

As within Audiovisual production, an important amount of information is generated during the production process; in past times most of it was discarded due to the difficulty in keeping incoherent documents related to each step of the process and not evidently necessary for the existence of the work. In the past also, due to the difficulty in replaying or remixing sequences, the results were often a “lucky shot” obtained after hours of essays with little or no possibilities of further remixing or major future modifications.

Today with the digitization of all the compositional process, new uses and necessities appear that were unthinkable years ago and which are often imposed by the evolution of uses and technology. Musical works as a “closed” item is more a consequence of the production constraints than of the desires of composers. More and more emphasis is being put in the last years in the performance of Acousmatic Music in a concert situation; performance determines the way a music is presented to the listener: how the space is organised and what loudspeaker configuration is used.

These considerations permit to identify different levels of preservation needs:

1. **Preserving the musical integrity of the work:** This is the main issue; preserving the information that determines what a work is. In the past, each work was associated to a physical media (tape, DAT, CD, etc.), which seemed to give certain integrity to the work and the way preservation concentrated in preserving the object. Today it is clearly considered that an Acousmatic work is content, with the same preservation issues as any other digital content that needs to be migrated regularly, to be checked in its digital coherence and integrity.

However, the main question remains: *How do I know that what I am listening to is the expected result?* This is why it is necessary to keep during the production process the elements that lead to the final work, as well as a complete description of the processes applied to sounds. This may permit a comparison and eventually a reconstruction of lacking or lost information. At the end of the work, the author should make a spectral analysis of the work; this will permit in the future to compare the signal as it is heard and the signal as it appears in a sonogram as well as providing a visual description of the sound components.

It is important to remark a main difference between an Acousmatic work and other recordings, including traditional music: since the sound environment is different in each work, the lack of information (as it has happened in the past with tape recorded material that progressively loses the high frequencies) can have a strong incidence in the comprehension of the music. When high or low frequencies are lost in speech or instrumental music, this does not have a strong incidence in the comprehension of the information, due to the existence of a codified language and codified sounds known by the listener.

2. **Preserving the final mix of the work and the intermediate steps:** The final work will be digital and will be transferred to a media for concert performance or publishing. The Sound Sequencer Environment, in which the work was assembled, remains the main source for any modification or evolution of a work. It is the virtual representation of an assembly of tens to hundreds of sounds containing all the necessary details of how the elements were assembled. All the actions done by the composer on the sound are visible in this kind of procedural score.

It is becoming increasingly important to maintain this environment because of the evolution of uses:

- a. **Versioning:** More and more composers are asked to do different versions of a work, not only in length, but also in terms of frequency distribution for reverberation (a radio version is different than the concert version or a CD version, or an on-line version). In a more general approach: works are modified with time. The composer himself mainly does the versioning.





- b. **Analysis:** Music exists by the way consumers appropriate themselves of the musical work and explore or analyse it. Traditional scores were easy to access; contemporary Acousmatic Music is extremely difficult to study in detail, simply for consumer curiosity, or for musicological issues. The Sound Sequencer Environment is a unique access to composer's ideas and acts.
- c. **Following the composer's actions:** It is possible in most Sequencers to have an "undo" function and an "undo list" which permits to access all the different actions done by the author since the beginning of the work. The difficulty is to have a register with the main actions and not the details; this would permit to follow the evolution of his actions and decisions

The main difficulty concerning the preservation of Mixes done on Sound Sequencers is the permanent evolution and changes on these platforms. There is a specifically conceived interchange format developed by Avid, a technology provider, called "Open Media Framework" (OMF) Interchange, now a standard format for the interchange of digital media data among different platforms, the OMF format encapsulates all the information required to transport a variety of digital media such as audio, video, graphics, and still images, as well as the rules for combining and presenting the media. The format includes rules for identifying the original sources of the digital media data, and it can encapsulate both compressed and uncompressed digital media data. The OMF format is far from being universal, but many Sequence developers accept to import and export in this format.

3. **Preserving the performance information:** Acousmatic works exist through their performance in a concert situation. This event generally precedes any other use of the work, as radio programming, Internet publishing, CD editions, etc. The specific environment of the concert is a unique and extremely difficult to preserve situation, and one of the most significant in the comprehension of how the composer proposes his music to listening. Diagrams, photos or any description item are very useful to understand the performance of the work; these elements are practically the only practical way to record the experience since no system exists capable of describing the actual distribution of Sound in Space.
4. **Preserving the information generated during the creation process:** This mainly implies all the existing documentation that a composer has to create in order to organize his production material. Several description methodologies exist using different description formats of sound; these are:
  - a. Taxonomical description (describing the perceived possible cause of a sound)
  - b. Morphological description (describing the spectral characteristics of the sound and its behaviour through time)
  - c. Subjective description (describing the impression the sound produces in the listener)
  - d. Technical description (describing the technology and the processes)
  - e. No specific description environment exists today that would permit to integrate the annotations often done by hand. This has a relation with the way sound files are used in Sequencer environments, where they are not encapsulated so it is practically impossible to attach written information.
5. **Preserving related material:** The Acousmatic Musical work, its production Mix, its related material, and its performance information are completed by related material of different origins and formats:
  - a. Information about the author and the work (bios and program notes); mainly text
  - b. Related information needed for the performance; (scores, specific diagrams, performance instructions, etc.) these can be text or images
  - c. Photography; images of the composer, the situation, the production or any other photographic information





- d. Philosophical writings on the intentions, feelings or other general considerations about the author, his esthetical ideas, and the music itself.
- e. External material; the work may generate other writings (newspaper clips, programs, critics, articles, films, concert recordings) which are complementary information to be preserved with the work.

In certain cases, when the technological environment is in real-time using complex control systems, the preservation issues will be very similar to those applied by Ircam in preserving their works and should thus be followed.

**Conclusion:** The main issue for preserving Acousmatic works is very close to the preservation of works of art. The production scheme generates an important amount of non-coherent material that can be of great use to guarantee the work's integrity. There is an increasing demand for complementary information regarding Acousmatic works as well as a need to access to the constitutive elements for further comprehension and description. Ideally, once a work is finished, a complete description of all the consecutive actions done by the author in order to obtain the result should be enough to preserve the work that could then be reconstructed in any new environment. This is very far from being possible today; this is why so many different information needs to be preserved.

The list of needed elements in order to assure a musical integrity of an Acousmatic musical work is then the following:

- **Preserving the musical integrity of the work**
- **Preserving the final mix of the work and the intermediate steps**
- **Preserving the performance information**
- **Preserving the information generated during the creation process**
- **Preserving related material**





CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and
Ingest	<p>Three stages:</p> <p>Production-related information is gathered by the composer, mainly sound descriptions and working notes. Other related information may have been produced: diagrams, screen copies, patch descriptions, settings in processing software. The composer decides what should keep, storage of performance scores is compulsory</p> <p>The Production Manager makes backup copies of the Sound-files and Sequencer-files (Mixes), to be kept by the author and in GRM.</p> <p>The work is entered in the Acousmaline Server with all the necessary identification information concerning the author, the work and the media</p>	-
Preservation Description Information	Provenance: Fixity Reference Context	All items need to be addressed
Representation Information	The format of the preserved items is determined by the file extensions.	Important changes in software and Sound-file formats may happen, necessity for a strong OMF like format for describing final Mix Sequences in a secured form. CASPAR framework should allow users to define and





		update the representation information for their data objects. Consistency check may be required to ensure that new representation information is compatible and interoperable with existing representation information.
Annotation		
Packaging	The relationship amongst components is a very important issue when accessing a work. Textual descriptions are never attached to Sound-files and should be linked strongly.	AIP needs to be defined
Description Information		Description of packaging needed
Access	Access to archive and ability to download is ensured by an online web interface ( <a href="http://acousmaline.ina.fr">http://acousmaline.ina.fr</a> ) Ability to locate the information concerning a work by: - Author, Name of the work, Version number, Photos - Written information (articles, critics, program notes) Ability to download the entire work or parts of the work as well as related information (CV, notes, photos)	All related data should be easily retrievable.
Access control, including DRM	Access control is based on a simple login/password No DRM is implemented at present time.	Issue about longevity of usernames/passwords – how long are they valid. What happens after the death of the authorised person? DRM should be implemented when there are rights associated to any document
Higher-level knowledge	Unclear what higher level knowledge is expected from users, other than the ability to render the digital content.	Issue: is there an assumed level of educational or cultural background assumed and if so, can it be captured?
Virtualisation and representation information	Sound and video streams	Potential virtualisation: digital processing algorithms Issue: possibly hierarchy of virtualised images: from still





		images to video.
Storage and storage virtualisation	No specific storage is used at present time.	
Preservation orchestration	None	
Authenticity	- Examples of results (sound samples...) are stored with the archive in order to know if the result gained in the performance is correct towards the intentions of the composer the evolution of the production staff (including performers).	Issue about authenticity of works: - Stored audio excerpts for comparison - Description of the intentions of the composer - Physical description of the signal (Sonograms) - Tools for helping to evaluate the integrity of a file





### 4.2.3 Preservation scenarios

Concerning the Preservation of Acousmatic objects the main components are:

- **Preserving the musical integrity of the work**
- **Preserving the final mix of the work and the intermediate steps**
- **Preserving the performance information**
- **Preserving related material and the information generated during the creation process**

#### 4.2.3.1 Changes in hardware and software

##### Scenario 1: Preserving the musical integrity of the work

It is the essential mission for a Preservation system to keep the work as it is, in its integrity; however different problems can be encountered:

The audio format in which the work is recorded: formats change or evolve, even if the exchange between formats is operational, little information is available on the possible consequences of transcoding on sound quality.

Identifying and differentiating versions of the same work as well as the quality of the copy.

Ensuring that all the related information or equipment is described and always available (scores, specific equipment for the performance)

##### Scenario 2: Preserving the final mix of the work and the intermediate steps

One of the main challenges today, Sequencers evolve very quickly and updating is impossible after two versions of the same software. It is necessary to dispose of a common description format (OMF Interchange is a good candidate) as well as the possibility to stratify the steps followed by the composer.

There is also a difficulty in keeping hundreds of files related to a reference mixing frame, where the loss of links may make the mix unusable.

#### 4.2.3.2 Changes in environment

##### Scenario 3: Preserving the performance information

It is essential to have a complete description of how the work was performed, with topographical information about the concert hall, technical description of the equipment and its distribution, as well as the space organization of the sound-files. Equipment changes very quickly, and even if the main functions remain alike, the interaction between the performer and the system evolve with time.

##### Scenario 4: Preserving related material and the information generated during the creation process

Large amounts of information are generated before and during the production process as well as once the work is finished. It is important to keep all that information which presents itself in different but common formats (text, images, films) and closely related to the works so to be able to access the large array of related documentation.

#### 4.2.3.3 Changes in designated community

##### Change in user community

Use proposed methodology from section 2.2.3.3





## 4.3 THE UNIVERSITY OF LEEDS TESTBED

### 4.3.1 Introduction

The University of Leeds is focusing on the testbed with Interactive Multimedia Performing Arts (IMPA). IMPA is getting popular for of real-time musical performances, especially those using gesture control systems [1-9]. In this testbed, we consider the preservation of IMPA is not only dealing with the digital multimedia output of such a performance but also the whole performing procedure and process in which the output is created, such as the mapping from gestures to multimedia contents, the order of operations, so that the same performance can be recreated at a later time. The preserved contents related to a performance also need to be synchronized during the recreation process. Therefore, preservation of IMPA is a challenging issue.

The test case used for this testbed is the preservation of Interactive Multimedia Performances produced by MvM (Music via Motion) system. The MvM system produces music by capturing user motions. Its overall system architecture is described in Fig. 2. The system captures user motions using motion capture devices. The data is then processed and stored in a 3D format. The captured data represented in 3D format is then passed to Motion Analysis and Recognition component of the system for identification of the performer's motions. These motions are then mapped into music, by using a mapping strategy, with parameters provided through a GUI. The motion-music map is forwarded to the generation component which produces multimedia content. In the case of the MvM framework, the content includes MIDI, setting of the mapping, video, audio and graphical animation.

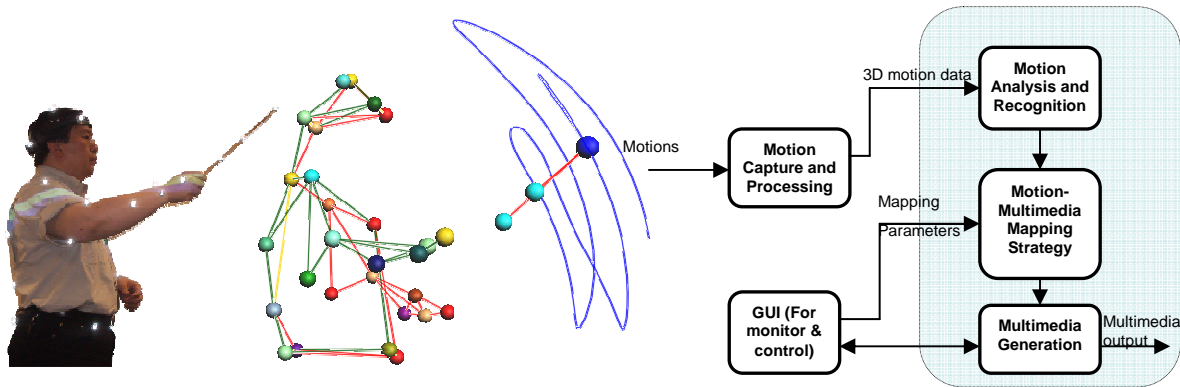
### 4.3.2 Preservation Scope

Creative interactive multimedia performances are generally *ad hoc*, in terms of system setting of mapping parameters as well as performers' gestures or motions. Therefore, it is very difficult to reproduce the same multimedia content at another time without preserving performers' movements and system setting.

For the MvM, the preservation required is at two different levels, for two different purposes:

- i. If only the output multimedia content and the performance itself are interested, the preservation will includes preserving still images, audio and video images captured during the performance. For analysis purpose, 3D motion captured during the performance also needs to be preserved.
- ii. For reproduction purpose (with or without performers' involvement), the preservation will cover the whole creation process. This involves preserving exact performers' creative gestures/motions (via captured motion data), mapping parameters and the software components that generate the music, including: Motion Analysis and Recognition, Mapping Strategy and Multimedia Generation. The dashed rectangle in Figure 1 shows the scope of MvM system that needs to be preserved.





**Figure 2: Preservation of interactive multimedia performances**

The preservation in case (ii) is complicated. It involves the preservation of both data and software. They require different approaches to preservation. For example, with data, this may involve the use of standard data formats; with software, relevant software components may be preserved. Typical multimedia environments used for such creative system include MAX/MSP [10], Pure Data (PD) [11]. In particular, to reconstruct the performance, the following elements may need to be preserved:

- Motions of the performer, captured and stored in 3D motion data files
- Performance scene setting, stored in 3D scene data files and video files
- The procedure in which the performance is set up and carried out
- A MAX/MSP patch (or other software) that contains the mapping strategy
- MAX/MSP application to run the patch
- The operating system on which the MAX/MSP application was running
- The hardware system used in the performance: the PC, cameras and their setting, amplifier, mixer and speakers.
- Finally, the music produced during the performance.

The preservation of IMPA needs to deal with the following challenging issues:

Firstly, in order to reconstruct an interactive multimedia performance, the whole production process, involving the performers' interactions with the multimedia systems, related hardware, needs to be preserved. Missing one of these components, the reconstruction process could be fail. The use of specialised software and hardware in interactive multimedia systems could complicate the problem, as any replacement of software and hardware may cause a loss to the integrity of a performance. Furthermore, the relationships amongst these components also need to be preserved, so that during the reconstruction process, these preserved components can be retrieved and assembled in the right way.

Secondly, preserving interactions between performers and the MvM system has a particular importance in preserving interactive performances. That leads to the need for a systematic way of modelling the captured interactions and archiving captured data. At the moment, a popular method to deal with this issue is to capture performers' motions and store them in 3D data format (for keeping exact position of performers in 3D space). However, there many different file formats currently exist for 3D motion data. They are usually specific to applications they born with. Majority of them are not formally and well documented. As there are many different formats, there are also many applications associated with the data formats. That makes the preservation process more difficult for which ever strategy is used.



### 4.3.3 Preservation issues

CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and
		<ul style="list-style-type: none"> <li>• <b>changes in hardware and environment (software, legal, social etc) and Designated Community</b></li> <li>• <b>loss of sources of Information (including loss of host archive)</b></li> </ul>
Ingest	The ingestion of data into archive is manually done by users.	CASPAR framework needs to allow users to submit a complete set of data objects required for preserving an interactive multimedia performance. As the relationship amongst components is essential during the reconstruction process, during ingesting, CASPAR framework needs to provide facility to capture that kind of relationships and to perform consistency check when accepting submitted data.
Preservation Description Information	Currently not in use	There can be different sets of data related to a performance, such as different MAX/MSP patches, different versions of 3D motion data. Each of these data sets can be used and modified by different users for a particular context. Therefore, managing provenance and usage contexts of different sets of data are necessary during the reconstruction process.
Representation Information	Representation information about the data is interpreted by the extension of the file in the archive. The extension of a data file is assigned automatically by the application that processes the data.	CASPAR framework should allow users to define and update the necessary representation information for their data objects. Consistency check may be required to ensure that new representation information is compatible and interoperable with existing representation information.
Annotation		
Packaging	All data related to a performance are stored in a folder,	AIP needs to be defined.





	including its subfolder. Interpretation of content within the folder is based on user knowledge	There is the need for a description mechanism to describe data related to a performance packed in a package in the archive. The description should state the relationship amongst data files in the package, there dependencies and their actual location in the package.
Description Information		Description needed
Access	Access to the archive is done manually, usually with the of the data owners (e.g. for the location of data, format of data in archived, etc.)	<p>Issue: need to capture knowledge of the data holders</p> <p>In a reversed direction of ingestion, the access function should allow users to retrieve all data objects required for reconstruction of a performance, using the relationship specified during ingestion. This goal can be achieved by using knowledge about a data package in archive captured from the data owners during its ingestion process. At the time of access, the knowledge of the designated user community about performing arts may be different from the knowledge the original data holders. Therefore, the mapping of knowledge of the original data holders about the archived package captured during its ingestion and the knowledge of designated user community at the time of access will also be necessary.</p> <p>The access function should also allow users to query the information about a particular performance, including date, time, performers involved and performance contents.</p>
Access control, including DRM	None	Currently, not necessary to this testbed, but would be a good feature to add.
Higher-level knowledge	Mostly based on the knowledge of the users of the data	General high-level knowledge about interactive multimedia performance in music will be necessary to provide a background understanding of a performance in archive. This general high-level knowledge will also be useful for preservation description of a performance, in relation to its external context.





Virtualisation and representation information	It is entirely dependent on the knowledge of data users for identifying and interpreting the data objects required for a performance and the relationship amongst them.	A performance object is complex data object, as explained earlier. It consists of a collection of individual data objects to be submitted to the archive and their relationships with external hardware. Therefore, the CASPAR framework needs to be able to provide description for representation information of individual data objects, the inter-relationships amongst data objects stored in archived and their relationships with external components (e.g. hardware)
Storage and storage virtualisation	None	Multimedia contents, such as video/audio files, are usually in large volume. Preservation of these contents would require large scale storages. A single storage will not be enough. Data grids for storing this kind of contents may necessary. Therefore, virtualisation of storage will be necessary
Preservation orchestration	None	A number of preservation strategies will be necessary for keeping the data objects about interactive multimedia performances usable in over time. Depending on the changes of in the future, emulation of hardware, migration of data or re-interpretation of archived data may be necessary.
Authenticity	None	Authenticity of a re-constructed performance is important when related archived data objects are migrated or emulated





#### 4.3.4 Preservation scenarios

This section describes a few possibilities that might happen during the preservation of an MvM performance. As discussed earlier, in order to preserve an MvM performance over time, preservation of the following components may be required.

- Motions of the performer, captured and stored in 3D motion data files
- Performance scene setting, stored in 3D scene data files and video files
- The procedure in which the performance is set up and carried out
- A MAX/MSP patch (or other software) that contains the mapping strategy
- MAX/MSP application to run the patch
- The operating system on which the MAX/MSP application was running
- The hardware system used in the performance: the PC, cameras and their setting, amplifier, mixer and speakers.
- Finally, the music produced during the performance.

The following are scenarios that may happen during the preservation process. They are in order of complexity.

##### 4.3.4.1 Changes in hardware and software

**Scenario 1: Required applications, the underlying operating systems and hardware are still in operation at the time of reconstruction.**

This is the easiest case. Only application data, such as 3D captured motion data, video/audio files, MAX/MSP patches are required to be preserved.

**Scenario 2: The underlying operating systems and hardware are still in operation at the time of reconstruction, but MAX/MSP and other related applications no longer exist.**

There are a number of possibilities for reconstructing a performance in this case.

Option 1: Preserving all the applications required for reconstruction of a performance in CASPAR storages.

Option 2: Although the applications used in the original performance are no longer exists, there may be applications with the same functions as those used to produce the original performance. If this is the case, the preservation only involves data objects as in the previous scenario. The preserved data can be fed to the available applications to reconstruct the performance. Data conversion from the preserved format to a new format required by new applications may be necessary.

**Scenario 3: The underlying hardware is still in place. However, the operating system on which the MAX/MSP relies has been changed.**

There are a few options:

Option 1: If there are applications with the same functionality exist in a new operating system. Then, the same procedure as in option 2 of scenario 2 can be applied.





**Option 2:** Preserving the operating system in CASPAR archive, then, applying the same procedure as in the previous scenarios.

**Option 3:** Migrating MAX/MSP and other related applications to a new operating system available at the time of reconstruction. This migration may involve redeveloping the software applications. The redevelopment may be costly.

**Option 4:** Building an emulator of the old operating to run on a new operating system. Then, the same operation as required for option 1 of scenario 2 can be applied.

#### **Scenario 4: The underlying operating systems and hardware have been changed.**

This is the case of very long term preservation. It requires a complicated procedure for preserving and reconstructing a performance.

- In the worst case, everything, from hardware, software and data, has to be preserved in order to successfully reconstruct a performance.
- Otherwise, all options in scenario 3 (except option 2) may be applicable.
- Hardware emulator may be possible. If this is the case, operations required for option 2 of scenario 3 can be applied.

The above description has shown various levels of complexity required for preservation of an interactive multimedia performance. Depending on the available hardware, software and our expectation of the situation in the future, appropriate action will be taken for preservation.

#### **4.3.4.2 Changes in environment**

##### **Scenario 5: Changes in organisations that owns the data objects in archive**

In an organisation, there are staff members responsible for maintaining the data in archive. A sudden change in these staff may lead to a situation in which no one in the organisation has access to the data stored in archive. This could be because of the lack of awareness or required information/knowledge to get control of the data. As a result, the data stored in the archive become obsolete. Over time, the obsolete data may become unusable, although they still kept in the archived. In order to avoid such a problem, it is necessary to well document the management of data stored in archive within the organisation that owns the data, so that substituting staffs can easily get hold of the control of data.

Another possibility is the change of organisational interests. In that case, although the organisation is aware of data packages stored in archive, however, they are no longer of interest of the organisation. Therefore, without necessary maintenance, the data packages in the archive will be out of data, sooner or later, and then become obsolete. In order to make use of these out-of-interest data, there should be a mechanism to transfer the management of data to other organisations who are interested in these data.

##### **Scenario 6: Changes in environment setting of a performance**

Setting of a performance is really important to interactive multimedia performances to maintain their authenticity. A preserved performance is more likely to be reconstructed in venues different from the original venue, where all the data in archived are captured. However, it is unlikely that the same physical setting of the venue could be recreated. Therefore, in order to maintain authenticity of a reconstructed performance, there is a need for a mechanism for describing physical setting a performance. This description should be





in a way that the setting will be able to be re-interpreted, where necessary, to adapt to the physical venues, where the re-construction will take place

#### **4.3.4.3 Changes in designated community**

##### **Change in user community**

Use proposed methodology from section 2.2.3.3





## 4.4 THE CIANT TESTBED

### 4.4.1 Introduction

CIANT maintains AMANT archive (<http://www.ama-nt.cz/>) aimed solely at video-art encoded in Real Video (combination of video and audio). AMANT's distinguishing features include DCMI compliance and the possibility for curators to evaluate and comment on the artworks. OASIS (<http://www.oasis-archive.info/>), which is being finalized these days, virtually combines number of distributed archives similar to AMANT maintained by other institutes by providing unified access to them. Existing databases are connected by a comprehensive meta archive system for artefacts like film, video, audio and image material in order to open up works of art historic value for the public. Plans exist to evolve from OASIS to GAMA by extending the range of archived data formats with 3D and other multimedia formats. These plans have been considered in this questionnaire.





#### 4.4.2 Preservation issues

CASPAR element	Summary of Current situation of testbed	Preservation issues arising from general considerations and <ul style="list-style-type: none"> <li>• changes in hardware and environment (software, legal, social etc) and Designated Community</li> <li>• loss of sources of Information (including loss of host archive)</li> </ul>									
Ingest	<p>Due to the nature of OASIS/GAMA project, origins of the data are unknown to some extent, since the content is provided by remote servers maintained by respective institute. Regarding AMANT archive that is under our control the ingest is a two step process:</p> <ol style="list-style-type: none"> <li>1. Manifestation(s) of the artwork is uploaded via WinSCP (secure file transfer protocol) to the media server connected to broadband Internet backbone.</li> <li>2. Description of the artwork (metadata) is filled and submitted using a web-based form together with URLs of the manifestation(s).</li> </ol> <p>The ingest happens only several times a week, typical artwork consists of one manifestation of an average size 100MB.</p>	[PAIMAS] checklist may be appropriate in improving preservation									
Preservation Description Information	<p>Reference:</p> <table border="1" data-bbox="394 1082 1400 1182"> <tr> <td>origin_id</td> <td>text</td> <td>OASIS Identifier unique in remote database</td> </tr> <tr> <td>identifier [Q]</td> <td>text[]</td> <td>Identifier, ISBN, ISSN, inventory number, etc.</td> </tr> </table> <p>Context information:</p> <table border="1" data-bbox="394 1230 1400 1305"> <tr> <td>related_objects [Q]</td> <td>text[]</td> <td>Links to related objects, uses OASIS Identifiers</td> </tr> </table>	origin_id	text	OASIS Identifier unique in remote database	identifier [Q]	text[]	Identifier, ISBN, ISSN, inventory number, etc.	related_objects [Q]	text[]	Links to related objects, uses OASIS Identifiers	<p>Reference:</p> <p>Issue: the longevity of the identifier and in particular an identifier involving databases has additional problems.</p> <p>Provenance:</p> <p>Fuller capture of processing history and custody needed.</p>
origin_id	text	OASIS Identifier unique in remote database									
identifier [Q]	text[]	Identifier, ISBN, ISSN, inventory number, etc.									
related_objects [Q]	text[]	Links to related objects, uses OASIS Identifiers									





	Fixity: N/A Provenance: <table border="1" data-bbox="394 331 1402 416"> <tr> <td data-bbox="394 331 667 416">provenance [Q]</td> <td data-bbox="667 331 813 416">text</td> <td data-bbox="813 331 1402 416">Determines which DB-Adapter (institute) provided the resource.</td> </tr> </table>	provenance [Q]	text	Determines which DB-Adapter (institute) provided the resource.	Fixity: needs to be addressed. Context: very broad context needs to be explained.						
provenance [Q]	text	Determines which DB-Adapter (institute) provided the resource.									
Representation Information	<p>OASIS/GAMA is a publicly accessible archive on the Internet. There are detailed help pages guiding new users how to access (search and interpret) the archived information. PHP bulletin board package is used to log feedback from users and provide answers to their questions. We do not provide any face-to-face training.</p> <p>In the case of OASIS/GAMA, users extract video and audio together with additional information like: authors, year, description, or physical location of the original material to name the most important attributes. The preserved information is an artwork that can consist of several manifestations (video, audio, and 3D models).</p> <p>This is a list of software that can be used to interpret the information encoded using the respective encoding:</p> <ul style="list-style-type: none"> <li>• MP3, MPEG2, MPEG4, WMV: Media Classic Player</li> <li>• RM: Real Media Video</li> <li>• VRML, X3D: Xj3D</li> </ul> <p>To watch the video, they must have installed a respective software according to the codec that has been used to encode the video stream. The basic information how to play the video and install the appropriate software is provided on the help pages. The interpretation of the art-work video is left up to the user.</p>										
Annotation	Content-Based Indexing Extension automatically extracts keywords from the video. <table border="1" data-bbox="394 1091 1402 1297"> <tr> <td data-bbox="394 1091 667 1142">keywords [Q]</td> <td data-bbox="667 1091 813 1142">text[]</td> <td data-bbox="813 1091 1402 1142">Keywords found in IndexingDatabase</td> </tr> <tr> <td data-bbox="394 1142 667 1222">timeoffsets</td> <td data-bbox="667 1142 813 1222">text[]</td> <td data-bbox="813 1142 1402 1222">Times of a keyword appearance in milliseconds counted from the beginning of the movie</td> </tr> <tr> <td data-bbox="394 1222 667 1297">weight</td> <td data-bbox="667 1222 813 1297">integer</td> <td data-bbox="813 1222 1402 1297">Relevance level, higher value means better matching result. Results will be sorted</td> </tr> </table>	keywords [Q]	text[]	Keywords found in IndexingDatabase	timeoffsets	text[]	Times of a keyword appearance in milliseconds counted from the beginning of the movie	weight	integer	Relevance level, higher value means better matching result. Results will be sorted	The algorithm and meaning of keywords and their relevance must be captured.
keywords [Q]	text[]	Keywords found in IndexingDatabase									
timeoffsets	text[]	Times of a keyword appearance in milliseconds counted from the beginning of the movie									
weight	integer	Relevance level, higher value means better matching result. Results will be sorted									





			according to this value.	
Packaging	<p>Each artwork consists of a few artwork manifestations that reside in separate files. Metadata describing the artwork reside in a SQL database. The extracted metadata associated with the artwork could be considered as an extra file. Files and the SQL database reside on remote servers maintained by respective institute. Remote databases provide indexing mechanism, character recognition and face recognition. The engine stores information about keywords/offsets found in the content. The searchable snapshot of the remote databases is created once per day.</p>			The AIP needs to be defined.
Descriptive Information	title [Q]	text	A name given to the resource	
	alt_title [Q]	text	Alternative title as a substitute or an alternative to formal title of the resource	
	id_system [Q]	enum[]	Identification system describes value of the Identification attribute.	
	person [Q]	text[]	Person (artist, author, editor, corporation, institute)	
	date [Q]	text	The year or date when the resource was created or first published	
	date_from	datetime	For specifying a range	
	date_to	datetime	For specifying a range	
	country [Q]	text[]	Country or list of countries where the resource was created or produced, ISO-Set	
	language [Q]	enum[]	Language, the language of the intellectual content of the audio-visual or text, ISO-Set	
	content [Q]	text	Content, abstract, synopsis, full text, table of content	





genres [Q]	text[]	Genres, in the case the source is describable as net art, painting, installation, etc.
metakeywords	text[]	Manually edited keywords
thumbnail	text[]	URL of the thumbnail
media_type [Q]	enum[]	Media type of the resource (application, audio, image, message, model, multipart, text, video) based on MIME ( <a href="http://www.iana.org/assignments/media-types/">http://www.iana.org/assignments/media-types/</a> )
oml_id	text[]	Identifier of a manifestation (digital or physical) (oml_* stands for online media list)
oml_url	text[]	Download link (e.g. video stream)
oml_format [Q]	text[]	Format
oml_description [Q]	text[]	Description of the format
oml_quality [Q]	text[]	Quality
oml_size [Q]	text[]	Resolution
oml_bitrate [Q]	text[]	Bitrate
oml_quality_rating	integer[]	Quality rating for content-based search in order to select the best quality manifestation
provenance [Q]	text	Determines which DB-Adapter (institute) provided the resource.
dba_base_url	text	URL of DB-Adapter





	preservation [Q]	text	Preservation method	
Access	<p>First, the users connect to the OASIS via web based fronted using internet browser of their choice. Then, they fill the search form and browse the results. We provide a web-based user interface which helps users to locate records in the archive. The search simply compares metadata associated with each record against the metadata in the query. The subset of metadata used for querying is marked with [Q] symbol in the table below the question “Semantic Representation Information”.</p> <p>The middleware uses caching database which is filled on-demand while loading metadata from remote servers. The software that provides access to the archive consists of a set of <i>OasisDBAdapters</i> for each archive, <i>OasisFrontend</i>, and <i>OasisMiddleware</i>. All the components were programmed using combination of PHP, PERL, and shell batches. The text source code of the components can be preserved easily, components are not based on any compiled binaries.</p> <p>Encoded videos are delivered using HTTP protocol and they are bit copies of the originals residing on the remote servers. The access frequency has not been measured yet, but we expect no more than one hundred accesses per day.</p>			
Access control, including DRM	<p>Every <i>OasisDBAdapter</i> uses authentication information provided by <i>OasisFrontend</i> to filter the search results. Thus, only a restricted subset of the matching set is provided. Particular policy dealing with ‘who can access what’ is left up to each institute.</p> <p>As far as AMANT is concerned, we distinguish two different access rights to the archive:</p> <ol style="list-style-type: none"> <li>1. public anonymous access. All the artworks in the AMANT archive have a low quality and low resolutions video thumbnail that is manually created during the ingest process. All the artists agreed that we can publish these thumbnails without any restrictions.</li> <li>2. restricted access. Access to the high quality and high resolution digital manifestations of the artworks is granted only to registered individuals: students, teachers, researchers, artists, etc. For a person to become registered,</li> </ol>			<p>Issue: the relevance and maintenance requirements on the list of registered users.and their access rights. Here there is the additional complication of the actual access being controlled by s third party.</p>





	<p>she/he must sign an agreement stating that the use of the data is limited only to non for profit acts (research, education, etc.). CIANT keeps written and signed copyright transfer agreements with the respective authors that allow the limited use of their artworks.</p> <p>Since the OASIS is in fact only unified interface to number of different archives, the answer to this question is not simple. We are going to provide complex answer as soon as we receive response from the two remaining data holders (Center for Art and Media in Karlsruhe and Netherlands Media Art Institute in Amsterdam Montevideo).</p> <p>Our approach for AMANT has been described below the question “Are there any access restrictions?”. We have an general institution-author copyright agreement suited for the AMANT archive that allows us to:</p> <ol style="list-style-type: none"> <li>1. publish video thumbnails without any restrictions</li> <li>2. provide full videos to individuals for limited use only (non for profit, research, education)</li> </ol> <p>Currently, we keep a simple spreadsheet database of authors containing their contact information with links to written and signed agreements. Additionally, we also have a similar database for the registered individuals.</p> <p>While OASIS is an international project and it includes artworks of authors from non-European countries, the effective legal framework becomes a complex combination of national laws of the respective partners (Czech Republic, Germany, and Netherlands), EU law, and international law.</p> <table border="1" data-bbox="394 1062 1402 1145"> <tr> <td>rights [Q]</td> <td>text</td> <td>Rights for presentation, preservation, publication, lending, etc.</td> </tr> </table>	rights [Q]	text	Rights for presentation, preservation, publication, lending, etc.	
rights [Q]	text	Rights for presentation, preservation, publication, lending, etc.			
Higher-level knowledge		Significant higher level knowledge remains to be recognised and captured.			
Virtualisation and	<ul style="list-style-type: none"> <li>• MPEG2 [ISO/IEC 13818-1:2000] (video)</li> <li>• RM; Real Media Video [http://www.realnetworks.com/] (video)</li> </ul>	Could be virtualised audio and video streams.			





representation information	<ul style="list-style-type: none"> <li>• WMV; Windows Media Video [<a href="http://www.microsoft.com/">http://www.microsoft.com/</a>] (video)</li> <li>• VRML [ISO/IEC 14772-1:1997, ISO/IEC 14772-2:2004] (3D)</li> <li>• X3D [ISO/IEC 19775] (3D), X3D encodings: <ul style="list-style-type: none"> <li>○ XML and Classic VRM (ISO/IEC 19776:2005)</li> <li>○ Binary encoding (ISO/IEC FCD 19776-3)</li> </ul> </li> </ul> <p>Each artwork consists of a few artwork manifestations that reside in separate files. Metadata describing the artwork reside in a SQL database. The extracted metadata associated with the artwork could be considered as an extra file. Files and the SQL database reside on remote servers maintained by respective institute. Remote databases provide indexing mechanism, character recognition and face recognition. The engine stores information about keywords/offsets found in the content. The searchable snapshot of the remote databases is created once per day.</p> <ul style="list-style-type: none"> <li>• video, arbitrary resolution (usually from 320x200 to 1920x1080), arbitrary framerate (usually 15 to 30 fps)</li> <li>• audio</li> <li>• 3D models: vertices, edges, surfaces, materials</li> <li>• metadata: (30+) textual value attributes</li> </ul>	
Storage and storage virtualisation	All files are stored on remote server of each provider, there are no copies unless the provider implements mirroring. The physical media is usually a cluster of hard drives with random access.	
Preservation orchestration	N/A	
Authenticity	Not solved yet.	Issue: authenticity of data to be defined.





## 5 THE CULTURAL HERITAGE DOMAIN

### 5.1 THE WORLD HERITAGE SITE TESTBED (UNESCO)

#### 5.1.1 Introduction

Documenting cultural heritage sites has been a long tradition of the conservation community. However, to document such sites in digital form is a relatively new technology. During the last 15 years digital techniques for cultural heritage sites have been significantly evolving: photogrammetry became digital and although extremely expensive at the beginning, it is becoming now a days an affordable technique.

There is no a single place where all data for all cultural heritage sites is stored. Usually the data is spread all over. It is mainly at research organizations that for a specific reason are studying and measuring a specific site, when the data for that specific site is captured and stored.

At UNESCO the data referring to World Heritage sites corresponds to the minimum legal requirements that the countries must provide.

#### Sample data for the UNESCO's CASPAR testbeds:

UNESCO will provide data samples mainly representing two major users: *the UNESCO-audience* and *the conservation-community-audience*. These two audiences were selected for CASPAR because they do well represent the overall user community.

In order to fulfil the testbeds requirements for CASPAR, UNESCO will provide the following data samples:

- a) Data samples of the UNESCO data the inscription of World Heritage sites.
- b) UNESCO will, in close cooperation with UNESCO partners that are working in the area of documenting cultural heritage sites, provide good representative samples for CASPAR. These UNESCO partners are University of Rome (CNR-Italy), Federal Polytechnical School of Switzerland Department of Geodesy and Remote Sensing and University of Cape Town, Department of Geodesy.

#### Location of the testbed

In order to simplify the CASPAR testbed avoiding the need to work with data on several remote platforms, UNESCO will establish a unique and single computer platform, located in Paris, where all the samples will be stored and where all the testing of CASPAR software and associated functionalities will be done.

There will be only two major types of users:

#### I / UNESCO-audience.

This user represents all the countries to the World Heritage Convention. The user group deals mainly with the legal data necessary to present a cultural heritage site as a candidate to become a World Heritage site.

Since the data might come out of 191 countries members of UNESCO, UNESCO cannot force any standardization of the data, otherwise developing countries would be out of the game. Therefore the data comes in all types of formats, sometimes analog only and sometimes totally digital.





### **Associated data.**

All of the documentation and data on World Heritage cultural sites which is held at UNESCO premises represents the justification for the sites' inscription. This is official data which is needed within the context of an International Convention to provide a legal record of a successful inscription. This means that they have met all the requirements for nomination. The data contains the following:

- Identification of property
- Description of property
- Justification of inscription
- State of conservation and factors affecting the property
- Protection and Management
- Monitoring
- Documentation
- Contact information of responsible authorities
- Signature on behalf of the State Party(ies)

The following well established procedures might happen to such a file:

a) The candidate cultural heritage site is not accepted and the candidature is deferred. In this case the file remains alive and the file will receive any updates that the country will send in order to improve the quality of information of the file so that the cultural heritage site can be re-presented as a candidate on future official meetings of the Convention.

b) The candidate is accepted. In this case, the cultural heritage site changes status from 'candidate' to 'inscribed site' (e.g. inscribed as a World Heritage site). In this case the file starts a new life. The file will receive any updates that the country wants to add to it and/or that the Convention decides on the site: State of Conservation, Periodic Reporting, etc.

### **For the UNESCO-audience, samples of the following sites will be provided as samples for the CASPAR testbed:**

- Franciscan Missions in the Sierra Gorda of Queretaro, Mexico
- Virunga National Park, Democratic Rep. of Congo and Uganda
- Vizcaya Bridge, Spain

## **II Conservation-Community-Audience**

This user represents the national conservation authorities and/or research institutions and that are dealing with the conservation of cultural heritage sites. In order to do such a conservation activity, they require to document the cultural heritage site as complete as possible. This group does not represent a unique user. For each site there are a large number of multidisciplinary experts. Each expert is responsible to capture and to store, retrieve and manage the data under his/her expertise.

This makes that there is no single standard about which data needs to be collected and how much data needs to be collected.

A good example would be a geographical expert who deals with all aspects of cartography associated to a site. This is in reality a multi-scale cartography that can include:

- Large-scale cartography to represent the total cultural landscape on which the cultural site is located.
- Detailed scale cartography representing the cultural heritage site as a whole.
- Even more detailed cartography indicating each individual monument of the cultural heritage site.





All of the above is supported by a photogrammetry expert who deals with all aspects of capturing, storing, retrieving and visualizing all measurement points that enable the complete and detailed documentation of each individual façade for each individual monument. The resolution varies according to the complexity of the details of each façade: a plain façade has usually low resolution since few measuring points are required, while a complex façade full of ornaments graven on the stone will require extremely high-resolution and therefore an enormous amount of measurement points.

#### **Associated data.**

In order to support CASPAR, UNESCO will provide samples data from three major partners:

- University of Rome, Italian National Research Centre (CNR) with high expertise in archaeology.
- University of Cape Town, department of geodesy, with high experience in photogrammetry and who has been developing during the last eighth-years the African Heritage database. This database has approx. 10 cultural heritage sites from Africa.
- Federal Polytechnical School of Zurich (ETH-Zurich), with a high expertise in remote sensing and photogrammetry, who has more than 20 years experience documenting cultural heritage sites in order to be able to reproduce these sites virtually on a computer in 3D, virtual tours, etc.

#### **In general the overall data process is as follows:**

- A multidisciplinary set of experts captures, stores and retrieves cultural heritage site measurements.
- All data for the cultural heritage site is stored on a common computer platform with a specific directory per site, inside the directory there are sub-directories for the different types of data stored.
- It is important to notice that in the case of GIS (geographical information systems) the software creates a specific database of all the cartographic layers and associated attributes.
- Data is used out of computer platform. Sometimes using very specific software and/or hardware.

#### **For the conservation-community-audience, samples of the following sites will be provided as overall samples for the CASPAR testbed:**

- Labilela, Ethiopia
- Via Appia, Italy
- A location in Switzerland

### **5.1.2 Preservation significance**

#### **UNESCO**

##### **Importance**

All of the documentation and data on World Heritage cultural sites which is held at UNESCO premises represents the justification for the sites' inscription. This is official data which is needed within the context of an International Convention to provide a legal record of a successful inscription. This means that they have met all the requirements for nomination. The data contains the following:





1. Identification of property
2. Description of property
3. Justification of inscription
4. State of conservation and factors affecting the property
5. Protection and Management
6. Monitoring
7. Documentation
8. Contact information of responsible authorities
9. Signature on behalf of the State Party(ies)

**Uniqueness of data**

It is vital to preserve this official data as it can be used for legal purposes within the international framework of the World Heritage Convention.

**Uniqueness of holding**

UNESCO holds the official nomination records for 830 World Heritage sites (as per June 2006), as well as associated State of Conservation reports for selected sites.

**Size of user community**

The size of the user community for UNESCO is, as of April 2006, 182 countries as well as general public.

**Magnitude of data set**

There are currently 644 inscribed World Heritage cultural sites, 162 natural sites and 24 mixed sites (sites with both natural and cultural values). The magnitude of the data set received by UNESCO can differentiate between countries. As explained before there are minimum requirements (see (i)) however countries try to strengthen their nomination file by sending in addition a large series of books, videos, DVDs, etc.

**Urgency of preservation needs**

The preservation of data for UNESCO involves the preservation of data and certain software.

**Conservation community****Importance**

At site level, the data represents the status of the site at a fixed moment in time. Time continues to modify the site, there is a normal deterioration. The site data including measurements, cartography and images is the only available data to know the details of the conservation status of a site at a fixed moment in time.

**Uniqueness of data**

The best example is the Buddhas of Bamiyan: the Buddhas were exploded and they no longer exist in a complete state, therefore the data that was collected on the Buddhas before the explosion has become the unique data available to allow humanity to know the original appearance of the Buddhas. The data is valuable as the cultural heritage can now be recreated through 3D modelling. Losing this data means losing such knowledge forever.

**Uniqueness of holding**

The whole set of data objects constitutes the single and unique digital representation of all cultural heritage sites stored in the database. This Information must be preserved since it represents the only existing knowledge about all these cultural sites.

**Size of user community**



The size of the conservation community is a large audience. It includes national conservation authorities, research institutions, universities and non-governmental organizations.

**Magnitude of data set**

As an example, from the Department of Geomatics, Univ. of Cape Town, the African Heritage Database holds currently Information for 10 cultural heritage sites and has a size of approx. 500 GB.

**Urgency of preservation needs**

The preservation in case (ii) is complicated. It involves the preservation of both data and software. They require different approaches to preservation. For example, with data, this may involve the use of standard data formats; with software, relevant software components may be preserved. In particular there is a need to preserve integrity of physical storage, support for migration or emulation of obsolete hardware, software technologies, support for virtualization of elements, in order to make possible to reconstruct them from scratch. Due to importance of data it is important to preserve all data, software and hardware to re-create it.





### 5.1.3 Preservation issues

CASPAR element	Current situation of testbed	Preservation issues arising from general considerations and changes in hardware and environment (software, legal, social etc) and Designated Community loss of sources of Information (including loss of host archive)
Ingest	The ingestion of data into archive is manually done by users. Most of the data from the conservation community and UNESCO community requires proprietary software: to create virtual models, 3-D visualizations, Adobe, etc...	CASPAR should allow the entering of data: one-point-load. In addition data might require to be updated continuously: modification of existing data and/or entering new data objects
Preservation Description Information	Provenance Fixity Reference Context	All need to be captured
Representation Information	Representation information is interpreted by the extension of the file. Data is differentiated by the extension of the files with certain exceptions.	Issue: there are many proprietary formats involved which must be described or else the software preserved. CASPAR framework should allow users to define and update the necessary representation information for their data objects.
Annotation	None	





Packaging		Some formalised packaging techniques are necessary to tie together the various distributed sources of data. Issue: how to rely on distributed data sources – for example all data may have to be centralised
Description Information		
Access	Access to the archive is done manually. Our data is located in different points depending on communities involved (UNESCO or Conservation community).	CASPAR should allow us to access all data. Knowledge about data objects is essential since data must be extracted jointly with the corresponding meta-data.
Access control, including DRM	No access restrictions	
Higher-level knowledge	Mostly based on the knowledge of the users of the data.	CASPAR framework needs to capture how data objects are conceptually linked together. This also involves background information about UNESCO and World Heritage. Issue: what is the distinction between Context (PRI) and Representation Information?
Virtualisation and representation information	Images Map structures Time series	
Storage and storage virtualisation	3D glasses for observing 3D presentations	N/A
Preservation orchestration	Discussions are ongoing at ISPRS/CIPA [International Scientific Committee for Documentation and Architectural Photogrammetry] with respect to preserve digital heritage data but nothing concrete has emerged.	N/A





Authenticity	N/A	Issue: authenticity of data held at remote sites
--------------	-----	--





## 5.1.4 Preservation scenarios

### 5.1.4.1 Changes in hardware and software

#### Scenario 1: Obsolescence of data media

Data media for objects must be checked to assess if such a media is still within the market standards.

#### Scenario 2: Digital photogrammetry (aerial and close range) and associated meta-data

Data must be preserved, if a new image format appears, data must migrate to any new standard so that data can continue to be used with the new format. Application software might need to be adapted/changed.

#### Scenario 3: Computer visualization

This requires specific graphic cards. If such cards become obsolete, data must be re-formatted to continue working on new available graphic cards.

Application software might need to be adapted/changed.

#### Scenario 4: Computer platform standard hardware

If hardware becomes obsolete the cascade effect on the data and associated software that might not be able to run anymore must be analysed and corrected.

#### Scenario 5: Computer platform specific hardware: sophisticated graphic cards, etc.

If any special hardware device becomes obsolete data must be re-formatted to be able to continue working on the new hardware device. Software must also be adapted

#### Scenario 6: Computer platform software operating system

A change of operating system might bring a cascade effect mainly with the specific hardware and all the application software. Corrected actions must be taken.

Computer platform software application packages: GIS, remote sensing, visual tools and modeling, etc.

As described before any change in data format or hardware or operating system software will have an effect on the application software and associated corrective action must be taken.

### 5.1.4.2 Changes in environment

#### Scenario 7: Changes in Geographic information systems (GIS) and CAD and associated meta-data

Data must be preserved, if a new map format appears, data must migrate to any new standard so that data can continue to be used with the new format. Application software might need to be adapted/changed.

Data-media for this object must be checked to assess if such a media is still within the market standards

#### Scenario 8: Changes in Remote sensing and image processing and associated meta-data

Data must be preserved, if a new image format appears, data must migrate to any new standard so that data can continue to be used with the new format. Application software might need to be adapted/changed.

Data-media for this object must be checked to assess if such a media is still within the market standards





---

### **5.1.4.3 Changes in designated community**

#### **Change in user community**

Use proposed methodology from section 2.2.3.3

.





## 6 COMMON REQUIREMENTS

### 6.1 CHANGES IN HARDWARE AND SOFTWARE

Hardware/Software obsolescence affects the ability to run software capable of

1. reading file formats
2. processing data
3. rendering data
4. manipulating data.

We can see these issues manifesting themselves across domains. Some illustrative examples found in our testbeds are:

- NetCDF MST files from the scientific domain require specific software packages to permit files to be read
- MAX/MSP patches widely used in the artistic domain which are highly complex, unstable and have a high dependency on proprietary software
- GOME raw data from the scientific domain which requires conversion to the SAFE format
- Dependence on specialised graphics cards for 3D rendering of images in the cultural heritage domain

The testbeds have identified a clear requirement to provide solutions that permit the required information to survive the obsolescence of software and hardware currently used to read, render or work with the data. Ideally any solutions would avoid dependence on proprietary formats and software with their attendant preservation risks. However current cost and technology constraints mean that solutions provided by CASPAR for some scenarios will involve other preservation techniques. Solutions utilising software preservation, emulation and format conversion amongst other techniques will be employed in order to provide the realistic solutions that are urgently needed.

#### 6.1.1 Changes in storage technologies

All storage media have some inherent preservation risks due to the storage media decaying, and hardware/software obsolescence may affect retrieval of data from storage medium. Some cross-domain examples are:

- UNESCO's distributed archive requiring the management of a wide variety of storage media
- GOME raw data stored on DLT tapes
- IRCAM DAT tapes
- CCLRC data access copies on RAID with back up in the Atlas data store

The key common requirement would be the orchestration of advice to archive managers on the preservation issues surrounding the storage media they are currently using.

At a different level, fixity of data is a general issue across testbeds, even if not acknowledged explicitly.

### 6.2 CHANGES IN ENVIRONMENT

A range of organisations supply information within the present day designated user community of an archive. This information might not be incorporated into the archive, might not exist in digital form or not have been documented at all. This information is often essential for the discovery, comprehension and use of the archived data. The information which needs to be preserved falls into the following three categories:

1. *The knowledge the dataset is capable of imparting to the user*





The reasons why one may wish to use data within the archive (a description of the knowledge the archive was intended to preserve). Some cross domain examples of the origin of such information is as follows:

- Justifications for a site's inscription as a world heritage site
- Experimental proposals
- Curators' evaluation of artworks
- Performance descriptions

The core requirement is the capture of all such relevant information from an organisation and its ingestion into the archive. This requirement results from the fact that user awareness of the potential of information is vital for the reuse of any dataset. This information must be stored in such a way that a future access system would be capable of searching and retrieving the data which corresponds to the identified knowledge requirement.

Awareness of the knowledge that such data provides may also evolve over time due to changes in scientific knowledge or society. This is due the user community discovering more diverse types of knowledge which may be extracted from the data. The ability to add or annotate in some way additional information regarding the potential knowledge the data can provide is an additional requirement.

The terminology used in a domain might change over time, and this has implications for the interpretation of datasets.

## 2. *Data Provenance*

All testbeds exhibit the common requirement for the preservation of provenance information. This adds value to and in some cases is vital for the correct interpretation and use of the core data set, for example if it describes the processing chain employed to move from raw data to processed data. Some cross domain examples are:

- Information on authoring body, commissioning organisation, experimental scientist or group, architecture, ethnic group responsible for construction of architectural object
- Mechanisms for the data collection instruments, modes of operation and calibration.
- Human scaling techniques employed, recording techniques, artistic movements influencing the work, composers' intentions, and archaeological mapping techniques.

It is important to preserve and relate this provenance information to the corresponding data. It may also be important to supply some form of explanation as to the significance of this provenance information.

## 3. *How to use the retrieved data to extract knowledge*

All domain data sets exhibit the need for the preservation of representation information, the capacity to virtualise/render the data and information capable of regenerating the higher level knowledge which is required for the interpretation of the resulting digital objects.

Representation information examples

- Data dictionaries containing parameters
- Formal descriptions of file formats
- Thesauri

The capture of representation information in order to provide adequate description of key data entities is a key requirement. The relationship between data entities and any other information element within the archive needs to be captured and preserved.

Virtualisation/Rendering examples

- Software
- Digital processing algorithms





Capture of important processing and rendering software is a key requirement. It should be noted that any software that is captured will then be subject to the additional set of requirements arising for software.

Higher level information examples

- Textbooks, journals and reference materials
- Specifications for specialist musical instruments
- Instructional materials for correct use/operation of software
- Instructional materials for the use/operation of specialised instruments
- Instructional materials for correct interpretation or analysis of rendered data

The requirement is the capture of at least the minimal set of materials capable of reconstructing the knowledge base of the designated user community.

### 6.2.1 Changes in legal framework

All archives are potentially subject to evolving legal restrictions, be they at a governmental or organisational level (i.e. between an archive and a user community). A repository of information which describes the implications and changes in legislation relating to an archive holdings and any element of its designated community would be desirable.

Copyright restrictions on data, software, hardware and supporting information evolve over time and potentially expire. The result of this evolution is the need to ingest or release access to previously identified materials, information or data. Some form of monitoring of the identified copyrighted materials and owning institutions is required to facilitate this.

It should be noted that copyright changes do not occur at the same time for all related information for all end users, for example the copyright expires on a score/data from scientific experiment but supporting textbooks /journal articles or instructional manuals are still under copyright. This creates a common requirement across all data sets that access be dependent upon the unique combination of end user type and information unit (i.e. any component of a dissemination information packet).

### 6.3 CHANGES IN DESIGNATED COMMUNITY

In all the test cases discussed in this document the implicit and explicit changes in the knowledge base of the Designated Community provide the most difficult and varied types of preservation problems. The most significant of these are changes in the semantic knowledge base, for example terminology drift, but also encompassing the hardware/software and environment changes mentioned separately above.

It is worth noting that change in Designated Community, and in particular the implication for preservation of semantic content, is the type of change ignored in other preservation studies, despite being an immediate consequence of the widely accepted OAIS view of preservation. Of course for a piece of digitally encoded information the Designated Community is not independently defined – the definition comes from those claiming to preserve the information.

Although different in detail from discipline to discipline, common problems include:

- how can the Knowledge Base be defined and how can changes be tracked
- defining the scope of Representation Information – in particular that of Semantic Representation Information
- where data is processed, what is the relationship between Representation Information of data to the Provenance of data produced later in the process.





## 7 CONCLUSIONS

Detailed investigation of a number of datasets in various disciplines, following a structured questionnaire, has allowed us to identify a number of common preservation requirements as well as a number of discipline specific ones.

Using the OAIS Reference Model as the guide has focussed the questionnaire on preservation issues applicable across disciplines, including data as well as documents. Detailed consideration of specific datasets has also allowed us to identify significant issues with respect to the exact scope and definition OAIS concepts which deserve further work.

The OAIS Reference Model concept of preservation of digitally encoded information is based on the maintenance of the understandability and usability of that information. This view lies at the heart of the CASPAR external metrics and validation criteria [DoW]. Linking the scenarios to these will ensure that the testbeds will contribute to the validation of the CASPAR tools and procedures.

This document provides the first instalment of requirements and scenarios. A number of subsequent instalments will be provided through the project, widening the range of datasets and disciplines.





## REFERENCES

Reference	Details
DoW	CASPAR Description of Work ( <a href="http://www.casparpreserves.eu/Members/metaware/ReferenceDocuments/caspar-description-of-work/at_download/file">http://www.casparpreserves.eu/Members/metaware/ReferenceDocuments/caspar-description-of-work/at_download/file</a> )
EAST	ISO 15889, available at <a href="http://public.ccsds.org/publications/archive/644x0b2.pdf">http://public.ccsds.org/publications/archive/644x0b2.pdf</a> , tools available at <a href="http://east.cnes.fr">http://east.cnes.fr</a>
ERPANET-1	Erpanet project ( <a href="http://www.erpanet.org">http://www.erpanet.org</a> )
InterPARES-1	InterPARES project ( <a href="http://www.interpares.org">http://www.interpares.org</a> )
OAIS	Open Archival Information Systems Reference Model ( <a href="http://public.ccsds.org/publications/archive/650x0b1.pdf">http://public.ccsds.org/publications/archive/650x0b1.pdf</a> )
PAIMAS	Producer Archive Ingest Methodology Abstract Standard ( <a href="http://public.ccsds.org/publications/archive/651x0b1.pdf">http://public.ccsds.org/publications/archive/651x0b1.pdf</a> )
PROV-EU	EU Provenance project ( <a href="http://twiki.gridprovenance.org/bin/view/Provenance/WebHome">http://twiki.gridprovenance.org/bin/view/Provenance/WebHome</a> )
PROV-MYG	MyGrid Provenance project ( <a href="http://twiki.gridprovenance.org/bin/view/Provenance/WebHome">http://twiki.gridprovenance.org/bin/view/Provenance/WebHome</a> )
WARWICK-1	Warwick Workshop: Digital Curation and Preservation – Defining the research agenda for the next decade ( <a href="http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf">http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf</a> )
XFDU	<a href="http://sindbad.gsfc.nasa.gov/xfdu/">http://sindbad.gsfc.nasa.gov/xfdu/</a>
TFFPA	Task Force on the Permanent Access to the Records of Science – see <a href="http://tfpa.kb.nl/">http://tfpa.kb.nl/</a> for details of the Research programme and Strategy document.





## A1 OTHER COMMON REQUIREMENTS FROM THE WARWICK WORKSHOP

The following are taken from the Warwick Workshop [WARWICK-1].

### A1.1 COMMON RESEARCH ISSUES IDENTIFIED ACROSS ALL THREE DISCUSSION GROUPS

<b>Discovery and location</b>	1. Adopt or develop an agreed, <b>persistent, actionable, identifier for digital objects</b> , with associated name resolvers which are themselves persistent.
	2. Continue to develop <b>search and discovery tools</b> in partnership with relevant user groups.
	3. Develop more <b>detailed Data Models</b> for each domain and abstract out intra-domain and inter-domain commonalities.
<b>Trust</b>	4. Develop and integrate <b>DRM, provenance and authenticity</b> checking into ingest processes.
	5. Prototype and test national certification “badges” as <b>prototypes of certification processes</b> .
<b>Cost</b>	6. Continuing <b>data collection and modelling of costs</b> , with adequately complex parameterisation, over the life-cycle of different data types.
<b>Automation and Virtualisation</b>	7. Develop language to <b>describe data policy</b> demands and processes, together with associated support systems.
	8. Develop <b>collection oriented description</b> and transfer techniques.
	9. Develop data description tools and associated generic migration applications to <b>facilitate automation</b> .
	10. Develop <b>standardised intermediate forms</b> with sets of coder/decoder pairs to and from specific common formats.
	11. Develop <b>code generation tools</b> for automatically creating software for format migration.
	12. Develop techniques to allow <b>data virtualisation of common science objects</b> , with at least some discipline specific extensions.
	13. <b>Management and policy specifications</b> will be need to be formalised and virtualised.
	14. Further <b>virtualisation of knowledge</b> – including developments of interoperable and maintainable ontologies.
15. Develop <b>automatic processes for metadata extraction</b>	

### A1.2 SPECIFIC RESEARCH TOPICS

<b>Virtualisation</b>	16. Continuing work on ways of <b>describing information all the way from the bits upwards</b> , in standardised ways – “virtualisation”. Work is needed on each of the identified layers in section A1.2.
	17. <b>Knowledge virtualization</b> involving Ontologies and other Semantic Web developments are required to enable the characterization of the applicability of a set of relationships across a set of semantic terms.





	18. Develop use of <b>data format description languages</b> to characterize the structures present within a digital record, independently of the original creation application.
	19. It is important to make significant progress on dealing with <b>dynamic data including databases</b> , and object behaviour.
	20. <b>Representation Information tools</b> , probably via layers of virtualisation to allow appropriate normalisation, including mature tools for dealing with dynamic data including databases.
	21. Additional work on <b>preservation strategies and support tools</b> , from emulation to virtualisation.
	22. Develop increasingly powerful virtualisation <b>tools</b> and techniques, with a particular emphasis on knowledge technologies.
<b>Automation</b>	23. Develop protocols and information management exchange mechanisms, including synchronisation techniques for indices etc., to <b>support federations</b> .
	24. <b>Standardised APIs</b> for applications and data integration techniques
	25. Fuller development of <b>workflow systems and process definition</b> and control.
<b>Support</b>	26. Develop simple semantic <b>descriptions of Designated Communities</b> .
	27. <b>Standardise Registry/Repositories for Representation Information</b> to facilitate sharing.
	28. Develop <b>methodologies and services for archiving personal collections</b> of digital materials.
<b>Hardware</b>	29. Develop and <b>standardise interfaces</b> to allow “pluggable” storage hardware systems.
	30. <b>Standardise archive storage API</b> i.e. standardised storage virtualisation.
	31. Develop <b>certification processes for storage systems</b> .
	32. Undertake research to characterise types of read and <b>transmission errors</b> and the development of techniques which detect and potentially correct them.

### A1.3 POLICY AND INFRASTRUCTURE DEVELOPMENT

<b>National and international infrastructure</b>	33. The need for a national (and international) <b>roadmap for an infrastructure to support long-term curation and preservation</b> , which is underpinned by common policies and standards, that address the roles and responsibilities of the various stakeholder groups
	34. The need for a <b>significant increase in investment</b> at national and international level
	35. More support for <b>collaborative activities</b> , at national and international level, is needed
<b>Cross-cultural and cross-disciplinary</b>	36. A clearer understanding of the needs of diverse disciplines and <b>encouragement</b> for cross-disciplinary programmes is required
	37. More <b>training and accreditation</b> is required for information professionals
	38. More <b>advocacy is needed in support of changing the research culture</b> to embrace the challenges, and invest time up front, in the curation and preservation





	process
<b>Partnerships</b>	39. Work in partnership with <b>commercial system providers and with key interested parties</b> such as CERN and others, on error levels and developing affordable scalability
	40. Build <b>accredited community resources</b> , such as public data bases, with a cachet of contributing data.
	41. Work with data generating community, and their funders, to <b>encourage the adoption of standards</b> .
<b>Support Infrastructure</b>	42. A clearer understanding of information management with respect to legal issues such as <b>IPR and trust</b> is needed
	43. The need for <b>more best practice guidelines</b>
	44. Support for <b>persistent resolver services</b> for persistent identifiers
	45. Support for <b>shared registries/repositories</b> of various types of metadata, particularly Representation Information and curation tools
<b>Standards</b>	46. Progress <b>Ingest standards</b> (e.g. follow-on from PAIMAS standard).
	47. Progress <b>Audit and Certification</b> draft to full ISO standard.
	48. Set up <b>accreditation process</b> under the supervision of an international body.





## A2 THE PRE-QUESTIONNAIRE

Each repository is characterised at the outset by the following basic features.

1. Holdings: overview of the type of data held, and a list of data sets.
2. Data Set(s): For a selection of the datasets in the repository, offering a variety of challenges, provide a description of the digitally encoded information to be preserved, from the bit level to the knowledge it conveys to its user community. We do not at this stage need very detailed descriptions. In addition we need a brief descriptions of
  - any special importance, or special unique quality
  - access restrictions
  - what information/behaviour the data encodes
  - how the data is stored
  - how the required data is located and retrieved (including DRM and Legal issues)
  - what additional data, equipment or knowledge is employed to extract required information/behaviour from the data.
3. Data Producer: A brief description of the group, individual or institution that produced the data set.
4. User Community: A description of the current user community and the characteristics of the designated community for whom this data might be preserved.
5. Current preservation plans





## A3 FULL QUESTIONNAIRE

This questionnaire works best if the focus is on a single dataset - which may for example consist of one or more files. Broader collections, which consist of combinations of many simpler components, may also be covered but the descriptions will become very complex.

For each dataset, details are required of (guided by the CASPAR architecture):

- ingestion into the repository
- access control
- knowledge/behaviour encoded
- domain specific virtual objects e.g. sound recording, moving image, Earth observation image, Solar Terrestrial Physics dataset - these can be made from:
  - generic virtual objects e.g. images, tables, sequences, etc plus simple values
  - binary encodings of the information
  - storage mechanisms

and additional information about:

- production - description of the way in which the information is captured or created and how it gets to the repository
- current use:
  - finding aids
  - software used to access the digital encodings
  - software/mechanisms to use/perform the encoded information

For each of these broad headings we include some guidance of the intention and also some questions which are relevant. In addition some incomplete examples are provided using Science and Performing Arts data.

### **1 What Information/Performance/Behaviour does your current user(s) extract from this data and what needs preserving?**

This is quite an open ended question but it is essential to establish the nature of the information you are attempting to preserve as this helps to define what needs to be preserved. This definition does not limit the potential information the archive is capable of providing, but rather helps define a minimum level of information to be preserved.

#### **Examples**

##### **Science data**

- Electron density variation with height at the specified location and date/time. Combining data from different dates/times allows one to calculate the changes in electron density profile with time.

##### **Performing Arts data**

- Viewing the play - moving images with synchronised sound





- Details of actors, production team etc etc

## 2 What information do you provide to a new data user, and what support do you give them during their use of the data?

This is to ascertain if there is any useful information would be given to a new data user to help them get started using the data. More importantly, it is also intended to "get at" the types of information that are not written down but are typically asked for and are needed by users (a data FAQ) to produce results.

Inevitably there will always be information in the heads of the people that run the archive that is not written down, but would be useful to some users both now and in the future. This is also assuming that the people that created and run the archive are not around to help the "unborn users", so in future they will not have the support.

Typical Questions:

- Do you give out any training material that informs a new user about using the instruments/data/software?
- Do you provide any training days for new users that inform users about the instruments/data/software?
- Do you log support queries and answers?
- Can you think of any information that you or your colleagues know that would be useful to new users that is not written down?

## 3 A clear definition for the information contained in the dataset

The definition of the data set must be sufficient to clearly identify what information is being preserved and should if relevant include the authority or reasoning behind the assertion that the data is what the producers e.g. the particle physical data came from CERN a trusted provider or Video footage from a BBC archive. Any factors that could to affect the interpretation of the data and therefore the quality of information preserved should also be explored.

- What were the physical factors (e.g. hardware/instrumentation/recording equipment etc) involved in creating this data ? \*If any of physical factors in data creation e.g. calibration were found to be false would an additional information be required for data reconstruction and is this within the scope of your archive?
- What were the human factors involved in creating this data (interpretations/sociological factors/adopted schools of thought etc)? \*If any of human factors in data creation were found to be corrupt would additional information be required for data reconstruction and is this within the scope of your archive?
- What scientific/intellectual assumptions have been made during the data creation or gathering process that allow you to make the assertion that the data is what you say it is? \*If any the assumptions in data creation were found to be corrupt would additional information be required for data reconstruction and is this within the scope of your archive? \*Who will be responsible for monitoring these assumptions and be responsible for any system changes?
- What external metadata, digital or non-digital needs to be integrated with the base data?

### Examples

#### Science data





## Measurement of electron density in the upper atmosphere using EISCAT

1. What measurements were taken and when, and which units were these measured in?
2. Which instruments, and instrumental settings, were used to make these measurements?
3. What processing, if any, has been done on this data e.g calibrations of various kinds, instrumental signatures removed etc. What units are associated with the data.
4. How these instruments where calibrated?
5. What assumptions are made to allow you to correlate an instrument reading with a value for electron density?
6. Any model or hypothesis used to remove anomalies from readings?
7. Reputation of transmitting and recording station(s)?

## Performing Arts data

### Recording of a performance of Shakespeare play

1. What constitutes a performance?
2. What constitutes a record of a performance?
  1. Sound/Video recording
    1. Who or what organisation recorded the performance
    2. What techniques did they use
    3. What equipment was used 1 Critics Review (as an example of a record of a performance)
    4. Who paid/employed the critic
    5. What was known about the critic e.g. where they a known feminist, did they belong to a known political party
    6. Who was the intended audience or publisher of a review
  3. If sampled from a large set of information, how a representative sample of performances ensured?

## 4 How is the digitally encoded information ingested into the repository

- Where does it come from? How is this verified?
- How is it packaged?
- For one dataset, how many "files" does it consist of?
- Is the data transformed in any way?
- Is information added (e.g. additional metadata, references etc)?
- Data volume (of the particular collection and each granule, i.e. file, of the data) and the rate at which it arrives





## Examples

### Science data

#### STP examples

- The data is collected daily by FTP from remote sites - using agreed usernames and passwords.
- The data is checked for validity i.e. can be read by the appropriate software.
- The data collected is a single file which is the data collected in one day.
- The name of the file is made up from the date and start time of the data collection.
- The data is a text file, containing columns of text and numbers, which is renamed and placed in a directory which corresponds to calendar year (e.g. /xxxx/collection-name/2006/site-startdata-starttimezzz.data).
- The way in which the data and time are encoded in the filename is not documented but is defined in the ingest software. In the date and start time are in UTC.
- An additional file named "collection-name.chp" contains additional metadata including the names, types, units etc, of the columns in the text file. Repository access software has encoded within it the relationship between "collection-name.chp" and the data file.

### Performing Arts data

TBD

## 5 How is the required data currently located and accessed?

- What information do current users need/possess that allows them to locate the data which provides them with information that they are seeking?
- What search and retrieval software do they utilise in order to do this?
- What additional stored data does this software utilise?
- How will the designated community maintain awareness of the information's existence?
- Are there measures to preserve the access software and any hardware it depends on?
- Does the access software utilise any supplementary data, e.g. an index database, a thesaurus?
- Can the software that provides physical access to the data store be preserved?
- If the software/hardware were not preserved, would it be possible to perform manual search and retrieval? Is there sufficient documentation to enable this?
- Is the data used by the user a bit-copy of the data held in the archive or might it be created "on-the-fly"?
- How is the data packed for delivery to the user?
- What is the current rate at which this data is extracted from the repository?

## Examples

### Science data

#### Ionosphere Example:





The ability to search and retrieve data by time, co-ordinate or by variation pattern ( ie detect standard patterns of variation in thickness of ionosphere).

- the user can ask for data from an arbitrary time interval, even though the data is stored in files, each of which has the data for one whole day.
  - the repository software extracts data from the relevant files and creates a single output file ( in CDF 2.7 format using the Cluster-STP conventions, and this is embodied in the software). The file is emailed to the user.

### Performing Arts data

Shakespeare Performance example

Key word searching, use of proximity operators, use of standardised subject headings e.g. library of congress and thesaurus e.g. mapping of Othello to “the Moor”.

## 6 Are there any access restrictions?

- Are there any restrictions on whom or how you can access or use your data?
- What are there reasons for these restrictions?
- Who or what imposes them?
- Are these restrictions likely to change over time?
- What are your current security requirements?
- What or who needs to monitor evolving access rights or legal issues?
- Who will be responsible for any system changes to ensure correct access is maintained?

### Examples

#### Science data

Only individuals from countries which are members of the EISCAT consortium are allowed access to the data; this restriction is imposed by funding bodies.

#### Performing Arts data

Shakespeare Performance Example

Copyright restriction and cross border intellectual property concerns.

## 7 Identify common "domain objects" currently used

For example, special types of images, lists etc.

\*Can you provide a comprehensive listing and definition of all separate data entities (most granular type of data held within file) contained within the file? \*Can you fully describe any entity relationships?

- How does current software extract and instantiate these entities and their relationships
- Can you provide a complete list of data processing functions the software performs e.g. comparison of data objects?





- What hardware dependencies does this software have?
- Can this software be preserved?
- If the software were to become unusable, could its functions be reconstructed from the technical specifications of the file format, the data entity definitions and their relationships? If not, what further information would be required to do this?

## Examples

### Science data

#### Ionosphere Example

- visual 3 dimensional models of the Ionosphere which evolves over time
- tables of data which always have TIME or EPOCH as one of the columns
- multi-dimensional images where all the component 2-D images contain data from the same time and location but each one corresponds to a specific wavelength of light

### Performing Arts data

#### Performance for Shakespeare Example

- nothing discipline specific

## 8 Are these objects special cases of simpler objects?

For example is the special types of image made up of a "simple" image plus additional simple objects?

## Examples

### Science data

#### Ionosphere Example

- simple tables of data i.e. essentially columns of numbers and text
- 2-dimensional images

### Performing Arts data

#### Performance for Shakespeare Example

- images
- text
- moving images i.e. individual images which should be displayed sequentially after a specified interval of time, accompanied by a synchronised soundtrack.

## 9 What information is required to reconstruct the information objects or reproduce the performance or duplicate the required behaviour?

This aims at capturing fairly high level, general, Representation Information.

- If any factors assumed in the data creation were found to be erroneous, would additional information be required for data reconstruction, and is this within the scope of your archive? Such factors might be physical (e.g. calibration), human, or scientific (e.g. theoretical assumptions).





- Is anyone identified as responsible for monitoring these factors and initiating any changes required as a result?
- What external digital resources does the user refer to?
- What external non digital resources (i.e. books/microfilm) refer to?
- What external bodies/organisation does a user refer to?
- What is the knowledge base and skill set of current user that allows them to extract information from the data instance?
- How do your current users acquire the knowledge base and skill set which allows them to interpret the encoded information and what are these?
- What knowledge or skills gap might arise between the current user group and the designated user community?
- What effect would such a gap have upon the interpretation of the stored data, and on the ability to process it further?
- Is anyone identified as responsible for monitoring the community knowledge base and initiating changes as needed?
- Is there any software used in conjunction with the representation information? Can this representation software be preserved?
- If the representation software were to become unusable, could its functions be reconstructed from the technical specifications of the file format, the data entity definitions and their relationships? If not, what further information would be required to do this?
- What effect would the permanent loss of such representation have upon the interpretation of the stored data?

## Examples

### Science data

#### Ionosphere Example

- Co-ordinates system or maps.
- Text book on atmospheric physics
- British Meteorological Office Data
- Definition of data format and
- Extracted data entities might be lower boundary height, upper boundary height , for each these a relationship which maps them to following values for co-ordinates latitude longitude and time.

### Performing Arts data

#### Shakespeare Performance Example

- Contemporary Shakespeare texts
- First Folio Editions





- Prompt books
- Explanatory text on prompt book notation
- Relevant Journal Article
- Newspaper Articles from Time of performance
- Extracted entities may be HSB values for a pixel, pixel co-ordinates with frame resolution and frame rate which allow you to reconstruct a moving image.

## 10 Structure Representation Information – (non media dependent encoding)

Closely connected with single file formats, but also includes complex inter-related collections of files.

- Provide a list of file format(s) in which the data held
- What are the technical specifications of this/these file format(s)? The information derived from the technical specification should be sufficient to extract data values and assign them to the appropriate identifiable data entities from the binary data held in file (see TIFF technical specification below for example).  
<http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>
  - Byte Structure of field e.g. depth of header or how positions of data fields are indicated.
  - How field or data sections lengths are defined/tagged.
  - Encoding of information within fields e.g. ASCII and how the type of encoding for a specified field is identified.
  - Compression employed and sufficient information to decompress
  - Any integrity checking that may aid in file reconstruction if file partial damaged?
- Specify any packaging connecting various separate components together
- XML Schema - specifies structure of an XML file

### Examples

#### Science data

If the file format was Fits you could refer to the following site <http://fits.gsfc.nasa.gov/> which contains references to other sites containing the papers describing the various variations of FITS.

#### Shakespeare Performance Example

If the performance was recorded on the DVD in MPEG2 format <http://www.chiariglione.org/MPEG/standards/mpeg-2/mpeg-2.htm>

## 11 Semantic Representation Information – (tend to be non structure dependent)

- Provide a comprehensive listing and definition of all separate data entities (most granular type of data held within file) contained within the file





- Fully describe any entity relationships
- Additional information which accompanies an XML schema to explain the meaning of the tags and the semantic relationship between the elements of the XML file.

### Examples

#### Science data

- Data dictionary providing the meaning of keywords in a FITS file header
- Ontology connected with Solar Terrestrial Physics data
- paper defining the VOTable format

#### Performing Arts data

- Performing arts ontology?

## 12 How is the data physically stored?

- How many independent off-site copies are there?
- Is the data fundamentally random access or sequential access in nature?
- What is the physical media upon which the data is stored e.g. CD, SDLT tape? Is this likely to change in the next 5 years?
  - Can you provide any relevant technical specification and physical description of how this information is mechanically transferred onto the storage media?
- Was there any media specific encoding employed the data file was written to the physical media?
  - Can you provide decoding instruction which allow the file to be reconstructed?
  - Has any integrity checking mechanism been allowed which will assist in file reconstruction?
  - Is any metadata physically recorded along with the files e.g. time stamps or id of machine writing to the media
- What is the volume of data held?
- How will the integrity of the data store be maintained?
  - What is the expected bit-error rate (corrected and, if possible also the uncorrected rate)
- What disaster recovery procedures need to be put in place?
- What relevant daa policies are there in place, for example number of off-site copies, frequency of tape retensioning etc?
- What is the storage medium current lifetime?

### Examples

#### Science data





### **Ionosphere Example:**

- The “bits” of information may be serially recorded on to magnetic tape in the Atlas Data store in CCLRC.
- Individual files are combined into a "Virtual Tape"

### **Performing Arts data**

Shakespeare performance: The bits of information may be stored on a DVD (what sort?)

## **13 Are there any additional preservation requirements?**

- Are there any preservation requirements on the dataset imposed by governments, institutional policies etc?
- Does your institution have any specialist preservation or long term access requirement or standards that must be adhered to? If so please detail.
- If an archival copy of the data store was created would these restrictions still apply?

### **Persistent Preservation Infrastructure**

- Key store
- Registry
- Directory services
- Tools

### **Questions based on Components**

#### **Costs**

Costs for preserving the information.

#### **Existing tools**

**Existing systems: h/w and s/w**

**Info access requirements e.g. queries/browsing etc**

**Risks/obsolescence**

**Timescales**

- Funding horizon i.e. for how many years is the repository funded?

#### **Other**

**What are the objectives of preservation for you?**

in terms of : duration scope exploitation For artistic side, it can be reperformance of the work

**What do they do for preservation now?**

**What help would they like in their preservation effort?**

**Legal requirements**

**What do they want from “preservation”**

**Who are the users of the preserved information? What is their knowledge base?**

## **14 Digital Rights Management requirements**



**What is the current approach your organization is following concerning digital rights?**

This question aims to clarify what are the concrete actions your organization is currently taking in respect of digital rights. It is essential to verify what and how is recorded about the digital content you will preserve in Caspar.

Examples:

- Your organization maintains a list of rights available for any digital content item
- Your organization just keeps a list of authors
- There's a general institution-author copyright agreement

**What kind of copyright information are you currently keeping? What copyright-related information would you like to keep and preserve in the future?**

Examples:

- Copyright holder's personal data
- Contact information for rights clearance
- Copyright history
- Records for any transaction concerning the content
- Statement of rights available over the content for your institution
- Statement of rights over the content that your institution is allowed to transfer to third parties

**Do you currently keep an electronic record of digital rights?**

Example:

- The institution keeps a record as a spreadsheet / database
- The rights are embedded into the content physical or digital manifestations
- The rights are separately kept and linked to the content via identifiers

**Under which legal framework for copyright and licensing are you currently operating?**

Example: This may be common copyright law, international copyright laws, specific trade agreements, public content licenses, EU regulations, and so on.

**Is your organization the copyright owner of the content you are going to preserve or are you collecting content belonging to different copyright owners?**

Examples: Content is produced and collected directly by your organization. Therefore any usage is only subject on your organization policies and regulations.

**If your organization holds the rights for the content to be preserved, which are the conditions you're applying for public / restricted / personal access to such content?**

Example:

- The content is available for free to the general public, if an attribution of your copyright statement is kept in every copy of the content.
- Content is available under payment of a fee, or a subscription





- Content cannot be re-distributed
- Content is licensed under a Creative Commons licensing scheme

**Do you have any formal authorization from copyright holders for the distribution of digital content under their copyright? If so, do you have a specific agreement on terms for what you are allowed to distribute and for which purposes?**

Example: The copyright owner signs a form stating your organization is allowed to distribute their content for a fee under the payment of a royalty

**Do you have any formal authorization from copyright holders for the preservation of digital content under their copyright? Is content adaptation or transformation for preservation purposes allowed under this agreement?**

Example: The copyright owner signs a form stating your organization is free to do whatever is necessary to preserve their content

**How is your organization planning to give access to the preserved content? Which cost model are you going to apply? Is there any institution-specific regulation about this?**

Example:

- Your organization policies state that you cannot produce any profit from the content
- There's a fixed pricing model, to cover preservation costs
- An ad-hoc pricing scheme is used depending on the "customer" needs

**Do you perform rights clearance?**

Example: Do you currently ask permission to the copyright holders (within your institution, or external) every time you want to distribute content? (in an on-demand fashion)

**Is your content encrypted or using some kind of copy protection mechanism? Are you allowed and able to remove such mechanism?**

Example:

- Audio files may be encrypted using formats such as Windows Media, Mp3 pro, DVD-Audio
- A DVD contains encryption technology which does not allow copying or using on DVD players purchased in other countries. If you are the copyright holder such limitation may be removed using appropriate software.
- Content is available in source or raw format, so no encryption or copy protection mechanism is present

A self-contained version of the Questionnaire is available at the CASPAR public web site (see [http://www.casparpreserves.eu/Members/cclrc/ReferenceDocuments/caspar-test-case-questionnaire/at\\_download/file](http://www.casparpreserves.eu/Members/cclrc/ReferenceDocuments/caspar-test-case-questionnaire/at_download/file))

---

<sup>i</sup> <http://public.ccsds.org/publications/archive/650x0b1.pdf>

