Long term data and knowledge preservation for the Earth Sciences Archive

S. ALBANI (ESA)
D. Giaretta (STFC)
Outline

OAIS conformance
  Information Model
  Mandatory responsibilities
Preservation workflows
Key Components
Threats and counters to those threats
Typical scenarios
ESA and EO introduction

- ESA users worldwide have online access to ~5 PB of EO data
  - EO data provide global coverage of the Earth
  - Data volumes are increasing dramatically
  - Large requirements for accessing historical archives
- This unique dataset has to be preserved!
  - ESA is promoting a European EO LTDP Strategy
  - ESA is involved in several international preservation activities (PARSE.Insight, Alliance…)
- ESA has a complex distributed system architecture
  - based on the OAIS standard
  - providing producer oriented services for data archiving, data retrieval and processing management…
CASPAR Project

EU FP6 Integrated Project

Total spend approx. 16MEuro (8.8 MEuro from EU)

http://www.casparpreserves.eu
A conforming OAIS archive implementation shall support the model of information described in 2.2. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

A conforming OAIS archive shall fulfill the responsibilities listed in 3.1. Subsection 4 provides examples of the mechanisms that may be used to discharge the responsibilities identified in 3.1. These mechanisms are not required for conformance. It is expected that a separate standard, as noted in section 1.5, will be produced on which accreditation and certification processes can be built.
Preservation workflows

Package Description

Archival Information Package

Packaging Information

Content Information

Preservation Description Information

Data Object

Representation Information

Physical Object

Digital Object

Structure Information

Semantic Information

Other Representation Information

Reference Information

Provenance Information

Context Information

Fixity Information

Access Rights Information

Interpreted using

Interpreted using

Interpreted using

1

1

1...*
The Information Model is key

Recursion ends at KNOWLEDGE BASE of the DESIGNATED COMMUNITY

(this knowledge will change over time and region)
- Use application to find data in Repository
- Create DIP with enough RepInfo for the user (via DC profile)
- Obtain more RepInfo from Registry if necessary
Key Components
Modules and Dependencies: defining the Designated Community

- **README.txt**
- **TEXT EDITOR**
- **ENGLISH LANGUAGE**
- **WINDOWS XP**

**MULTIMEDIA PERFORMANCE DATA**
- **C3D**
- **DirectX**
- **MAX/MSP**
  - 3D motion data files
  - 3D scene data files
  - motion to music mapping strategy

- **FITs FILE**
- **FITs STANDARD**
  - PDF STANDARD
  - PDF s/w
- **FITs DICTIONARY**
  - FITs JAVA s/w
  - JAVA VM
  - UNICODE SPECIFICATION
  - XML SPECIFICATION
  - DICTIONARY SPECIFICATION
Modules and Dependencies: Examples
(Semantic Web data)
Scenario: Intelligibility-aware Packaging

- Gap(o2,P1) = ∅
- Gap(o2,P2) = {FITS, FITS_STANDARD, FITS_DICTIONARY, DICTIONARY_SPECIFICATION}
- Gap(o2,P3) = {FITS, FITS_STANDARD, FITS_DICTIONARY, DICTIONARY_SPECIFICATION, PDF_STANDARD, XML_SPECIFICATION, UNICODE_SPECIFICATION}
- Gap(o3,P3) = {ZIP}
- Gap(o3, ∅) = {ZIP, C3D, DirectX, MAX/MSP}
Creating an OAIS Archival Information Package
<table>
<thead>
<tr>
<th>Threat</th>
<th>Requirement for solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved</td>
<td></td>
</tr>
<tr>
<td>Non-maintainability of essential hardware, software or support environment may make the information inaccessible</td>
<td></td>
</tr>
<tr>
<td>The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity</td>
<td></td>
</tr>
<tr>
<td>Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future</td>
<td></td>
</tr>
<tr>
<td>Loss of ability to identify the location of data</td>
<td></td>
</tr>
<tr>
<td>The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future</td>
<td></td>
</tr>
<tr>
<td>The ones we trust to look after the digital holdings may let us down</td>
<td></td>
</tr>
</tbody>
</table>
Accelerated Lifetime tests

As part of the validation the CASPAR tested simulated the following:

- hardware changes
- software changes
- changes in the environment (including legal framework)
- changes to the knowledge bases of the Designated Communities
# Test scenarios vs Threats to digital preservation

<table>
<thead>
<tr>
<th>Threat</th>
<th>STFC</th>
<th>ESA</th>
<th>UNESCO</th>
<th>IRCAM</th>
<th>UnivLeeds</th>
<th>CIANT</th>
<th>INA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users may be unable to understand or use the data e.g. the semantics,</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>format, processes or algorithms involved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-maintainability of essential hardware, software or support</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>environment may make the information inaccessible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The chain of evidence may be lost and there may be lack of certainty</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>of provenance or authenticity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access and use restrictions may make it difficult to reuse data, or</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>alternatively may not be respected in future</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The current custodian of the data, whether an organisation or project,</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>may cease to exist at some point in the future</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
STFC Testbed – various STP data
The ESA selected dataset for the CASPAR scientific testbed consists of data from GOME (Global Ozone Monitoring Experiment), a sensor on board the ESA ERS-2 (European Remote Sensing) satellite.

Testbed Dataset

Preservation of the ability to process GOME data from L1B to L1C

MUF 21.95
M 3.570
D 3000

h'F 207
h'F2 N/A
h'E 100
h'Es 105

zmF2 213
zmF1 N/A
zmE 103
vF2 77
vF1 N/A
vE 14

C-level 51
The Villa Livia dataset is a collection of files used within the "virtual museum of the ancient Via Flaminia" project: a 3D reconstruction of several archaeological sites along the ancient Via Flaminia, the largest of them being Villa Livia.
This is an elevation grid (height map) of the area where Villa Liva is located. It is an ASCII file in the ESRI GRID file format.
Contemporary Art Testbed
Performance Viewer: side-by-side comparison and validation of the transformation. From left to right: 3D visualization in Ogre3D, 3D model of the stage including the virtual dancer in VRML.
Figure 8: Some aspects of acousmatic production.
CASPAR Validation

In all cases members of the Designated Community, with appropriate changes to mimic changes over time, verified that the metadata was adequate for the use despite simulated changes of hardware, software, environment and Designated Community over time.

Full details are available in the validation report (CASPAR Validation report, 2009)
CASPAR – http://www.casparpreserves.eu

CASPAR Source code -
http://sourceforge.net/projects/digitalpreserve/

OAIS Reference Model -
http://public.ccsds.org/publications/archive/650x0b1.pdf
and the updated draft is available from

CASPAR Validation report
http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-validation-evaluation-report/at_download/file