Long term data and knowledge preservation for the Earth Sciences Archive

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

PV 2009

Outline

OAIS conformance

- **Information Model**
- **Mandatory responsibilities**

Preservation workflows

Key Components

Threats and counters to those threats

Typical scenarios

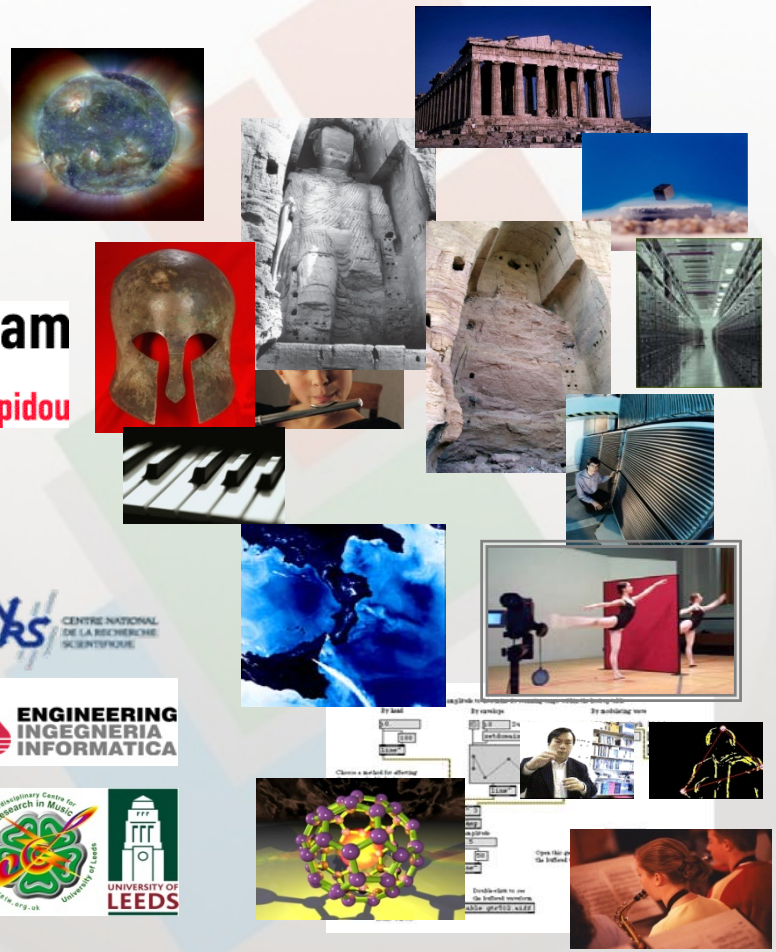
ESA and EO introduction

- 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)



Science & Technology Facilities Council

metaware

ircam Centre Pompidou

ACS ADVANCED COMPUTER SYSTEMS

UNESCO

UNIVERSITY of GLASGOW

IBM

CIARIT

@semantics worldwide information integration

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ina

FORTH ICS

ENGINEERING INGEGNERIA INFORMATICA

esa

Scientific Research in Music — UNIVERSITY OF LEEDS

http://www.casparpreserves.eu

Information Society and Media

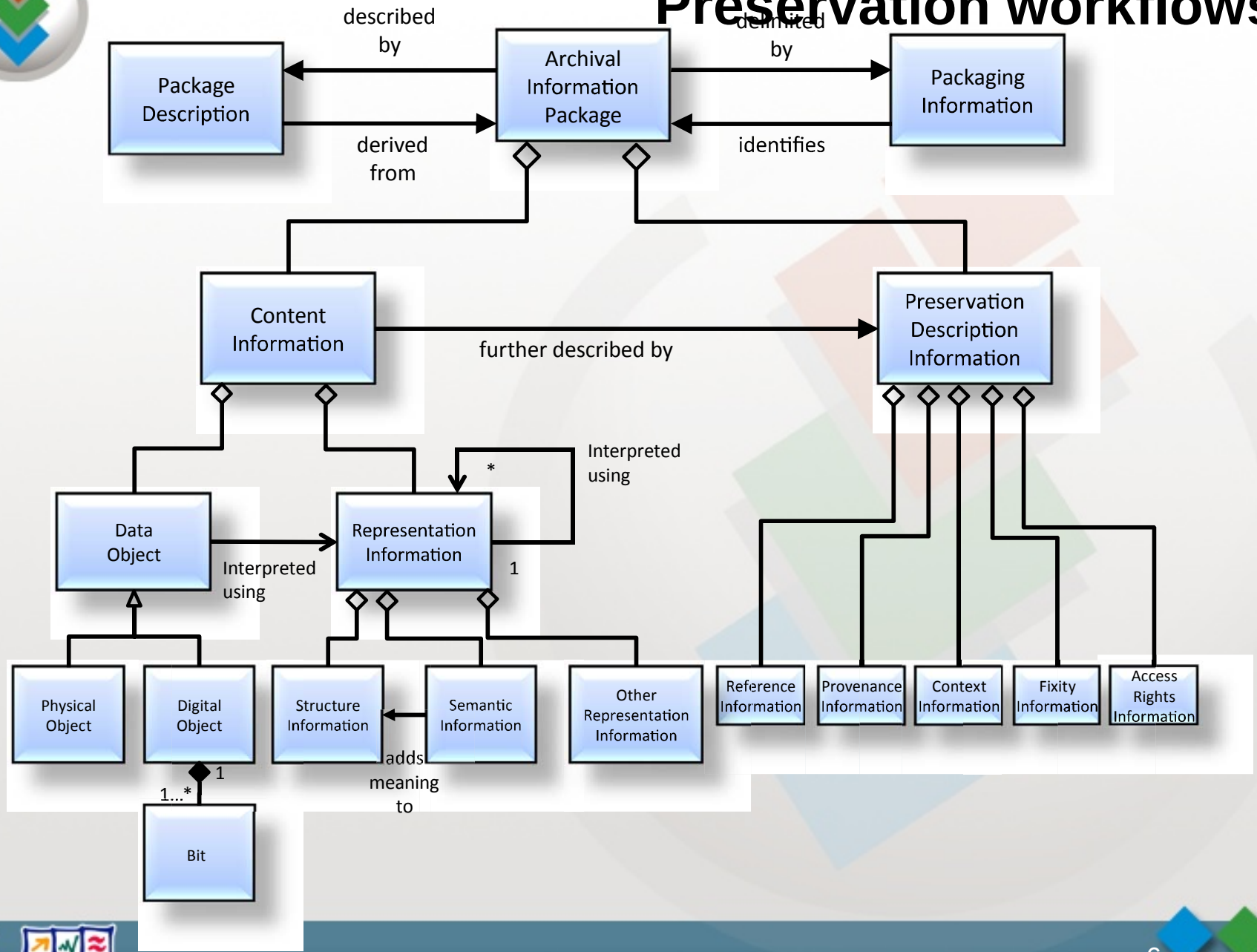
OAIS conformance

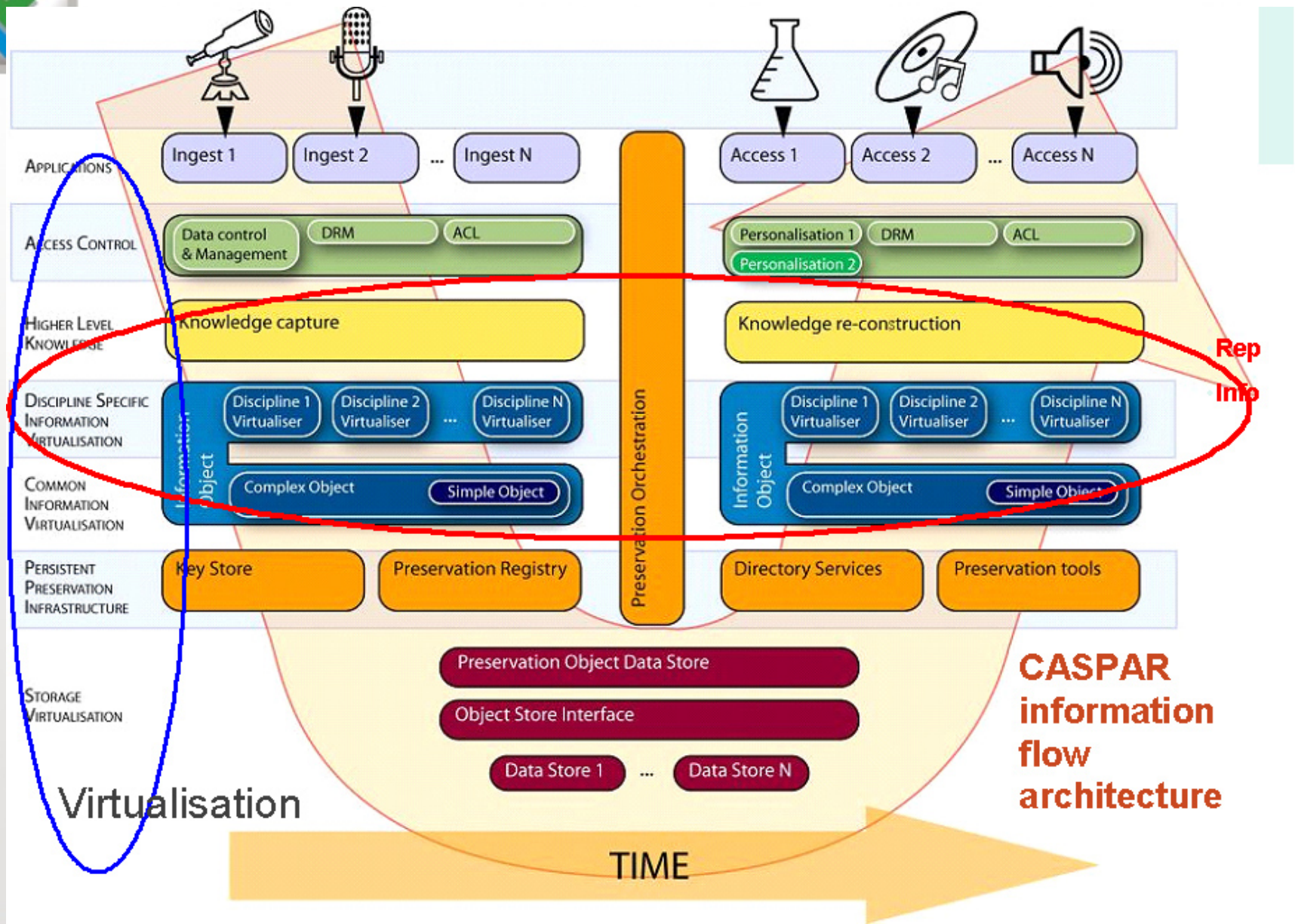
From OAIS:

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



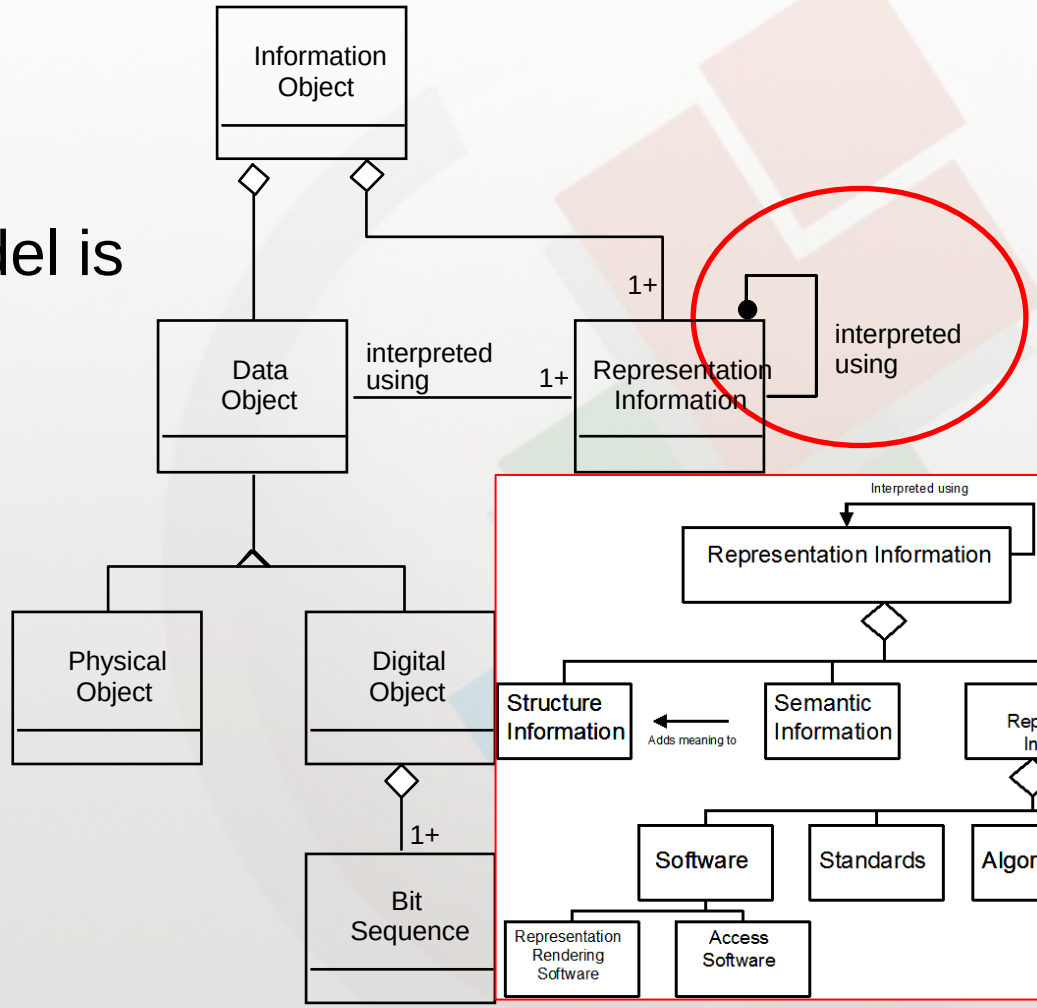

Ingest 1
Ingest 2
...
Ingest N
Access 1
Access 2
...
Access N
Applications
Access Control
Data control & Management
DRM
ACL
Personalisation 1
DRM
ACL
Personalisation 2
Higher Level Knowledge
Knowledge capture
Knowledge re-construction
Discipline Specific Information Virtualisation
Discipline 1 Virtualiser
Discipline 2 Virtualiser
...
Discipline N Virtualiser
Information Object
Common Information Virtualisation
Complex Object
Simple Object
Persistent Preservation Infrastructure
Key Store
Preservation Registry
Preservation Orchestration
Directory Services
Preservation tools
Storage Virtualisation
Preservation Object Data Store
Object Store Interface
Data Store 1
...
Data Store N
Virtualisation
TIME
Rep
Info
CASPAR information flow architecture

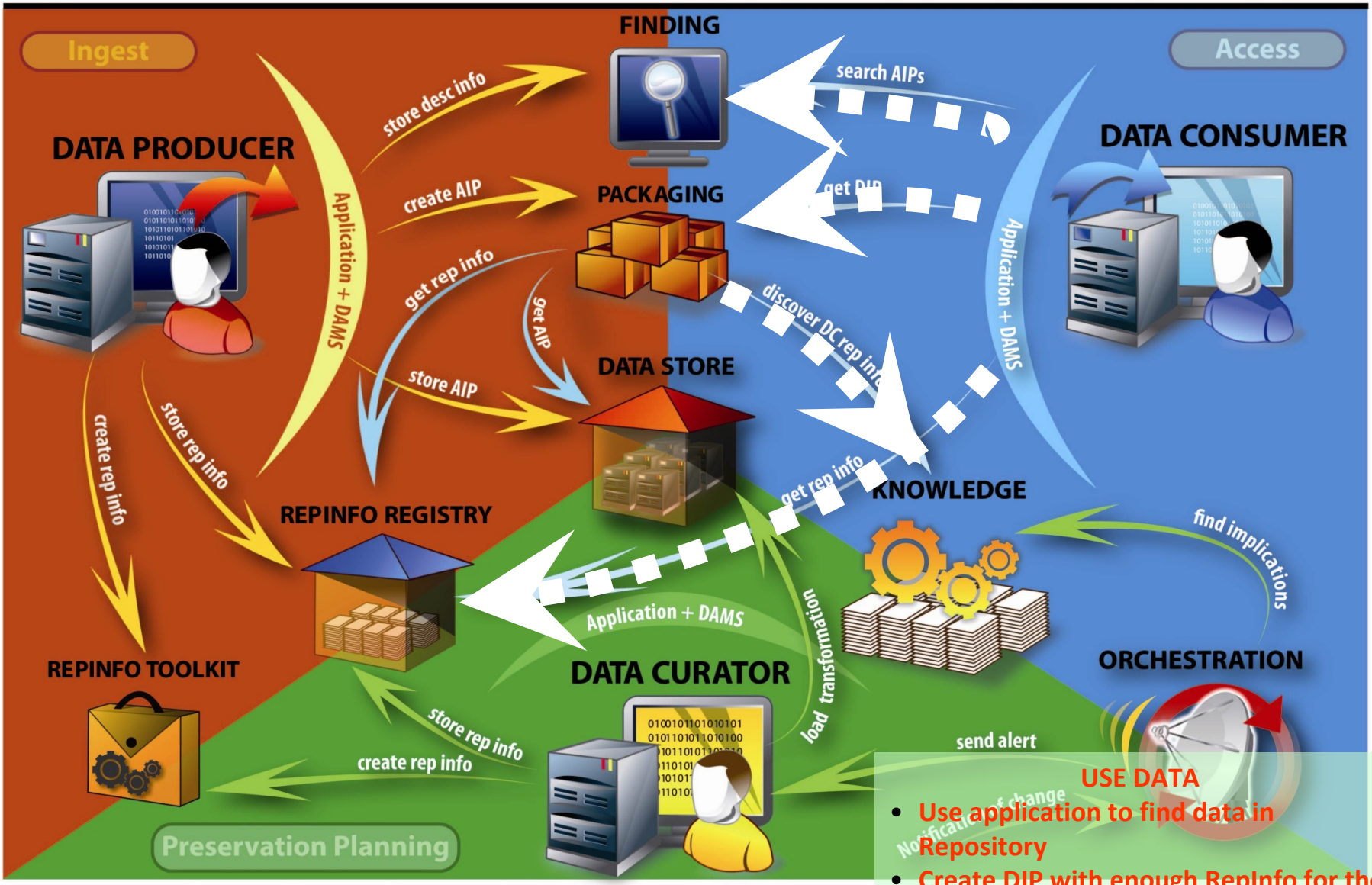
Information Model & Representation Information

The Information Model is key

Recursion ends at KNOWLEDGEBASE of the DESIGNATED COMMUNITY

(this knowledge will change over time and region)

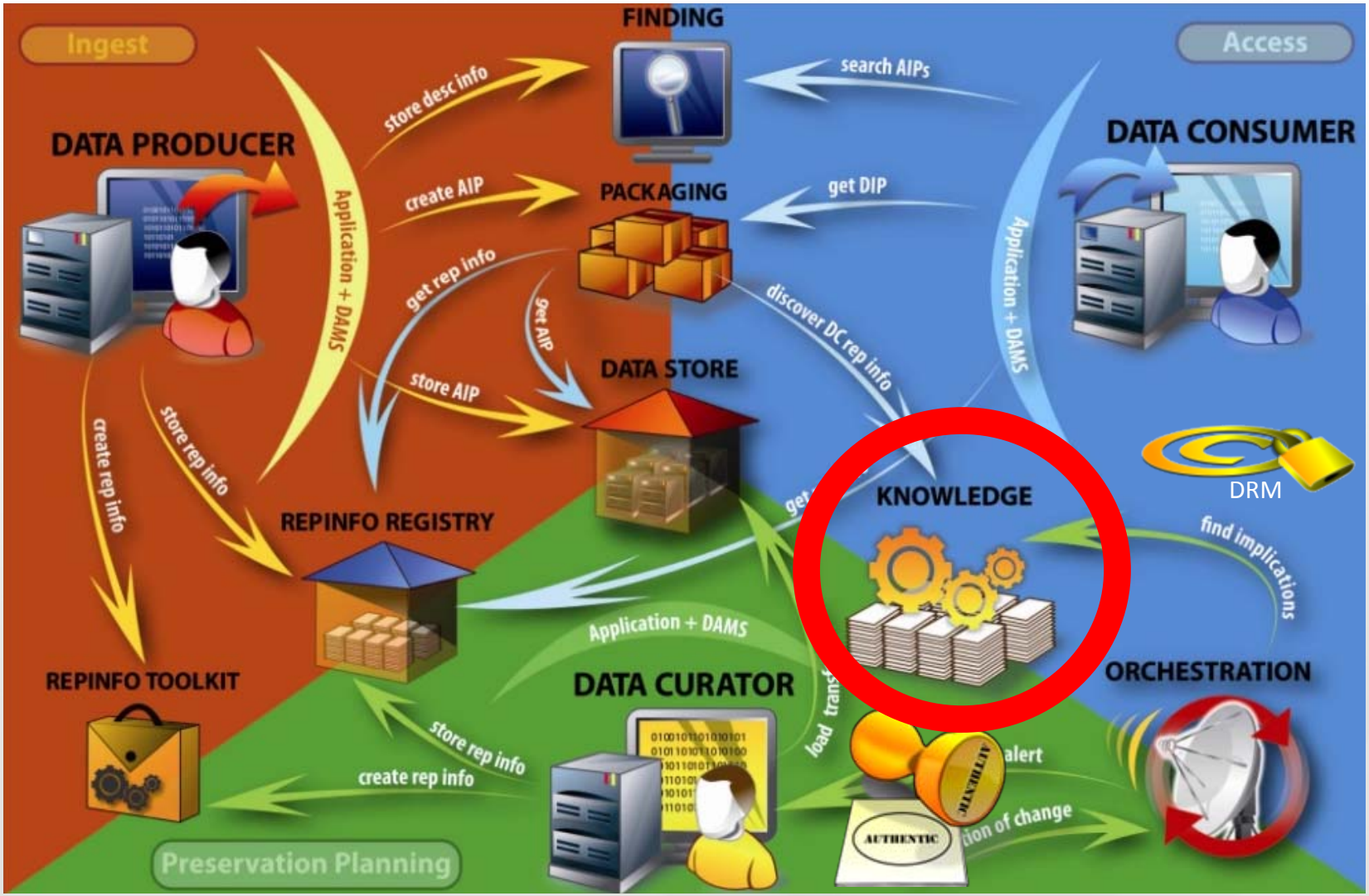




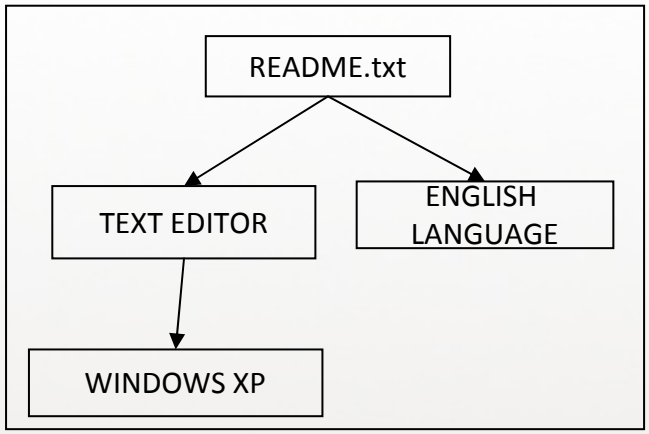
USE DATA

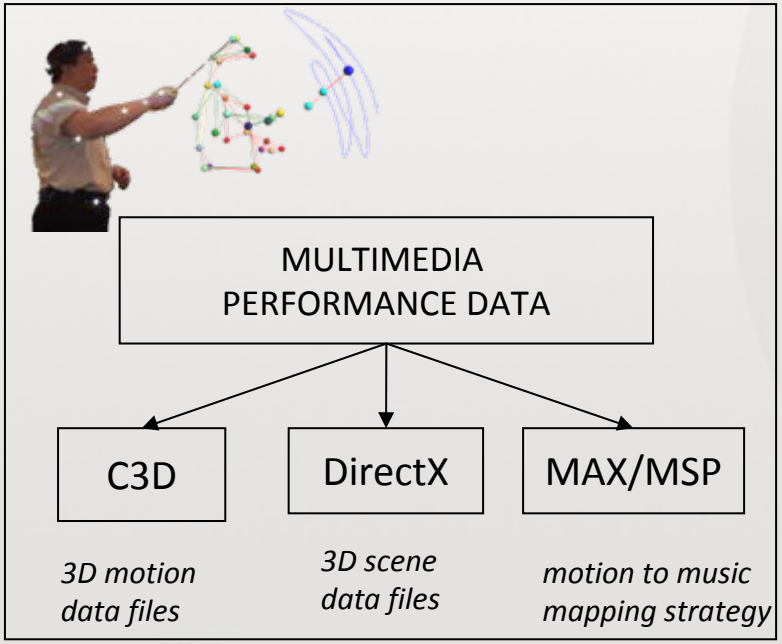
- Use application to find data in Repository
- Create DIP with enough RepInfo for the user (via DC profile)
- Obtain more RepInfo from Registry if necessary

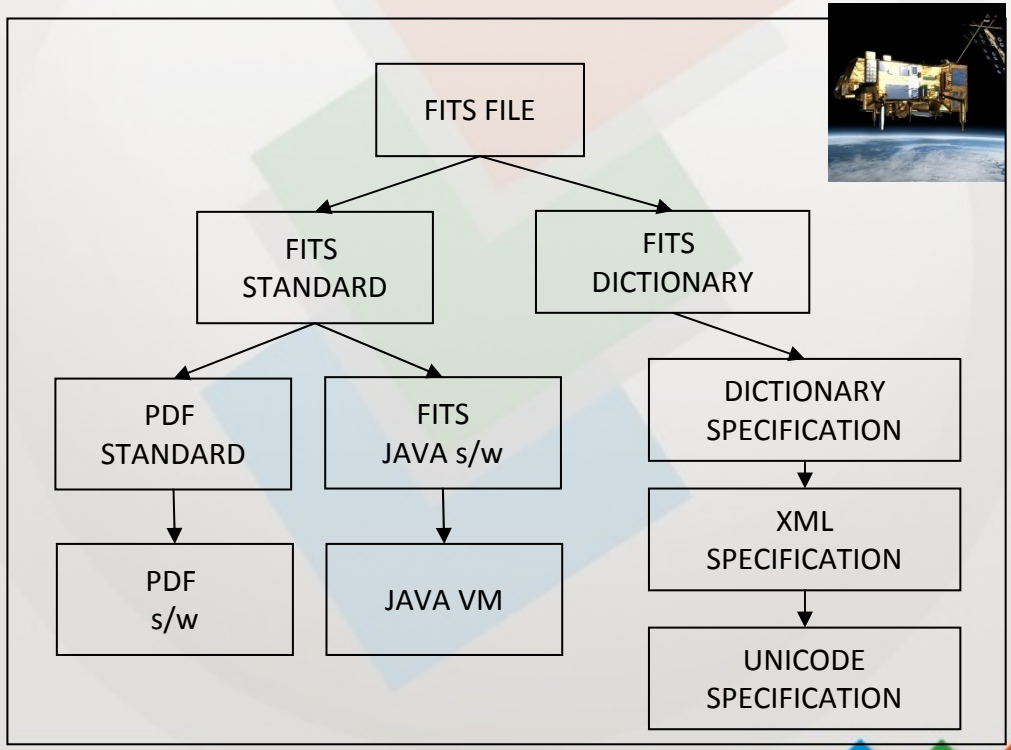
Key Components



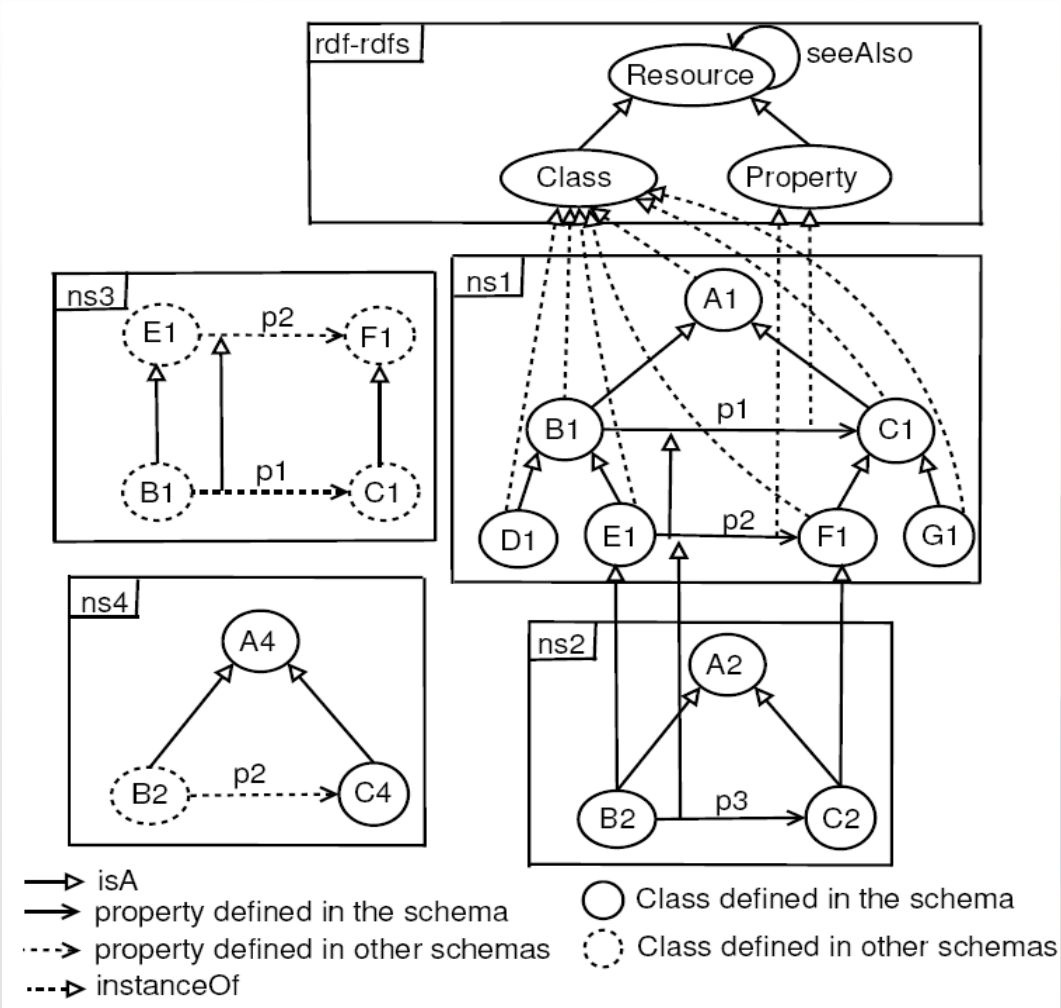
Modules and Dependencies: defining the Designated Community



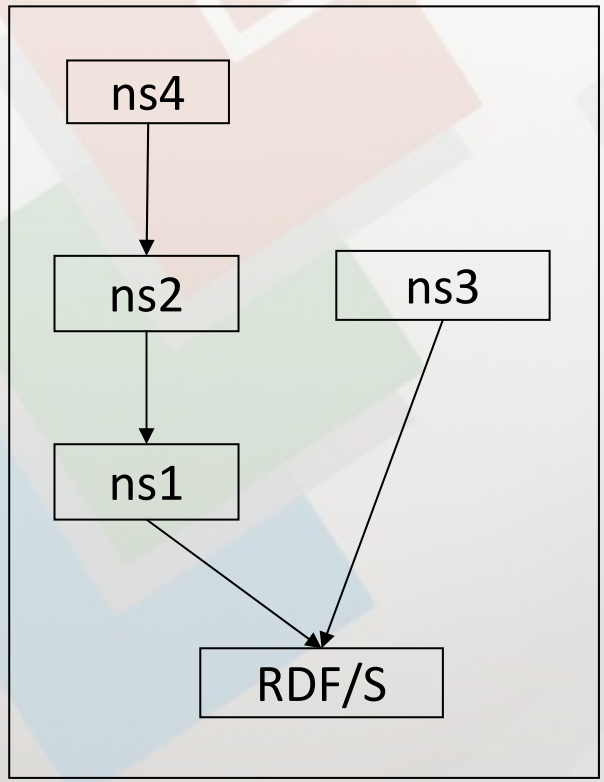




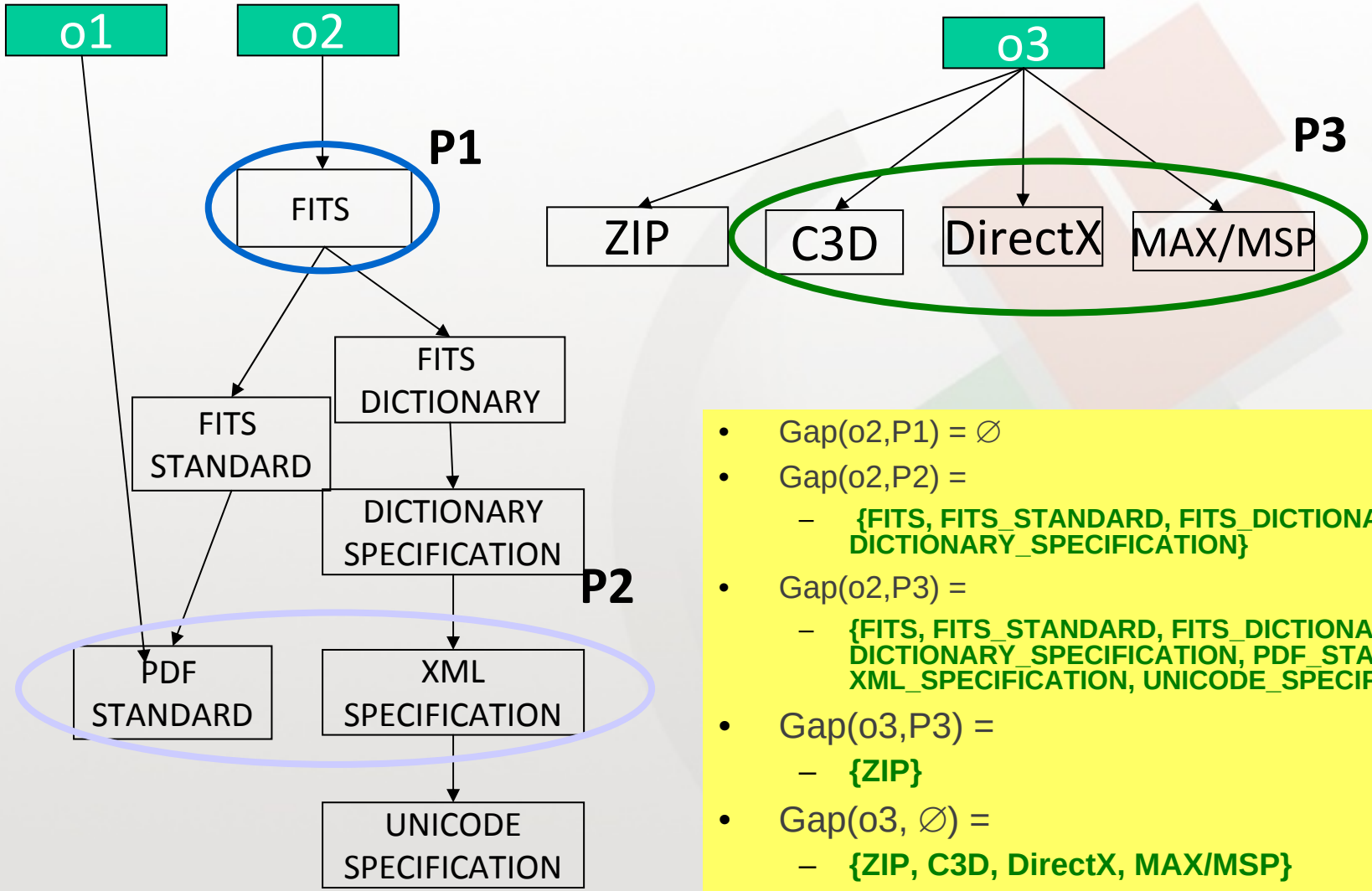
Modules and Dependencies: Examples (Semantic Web data)



modules and dependencies

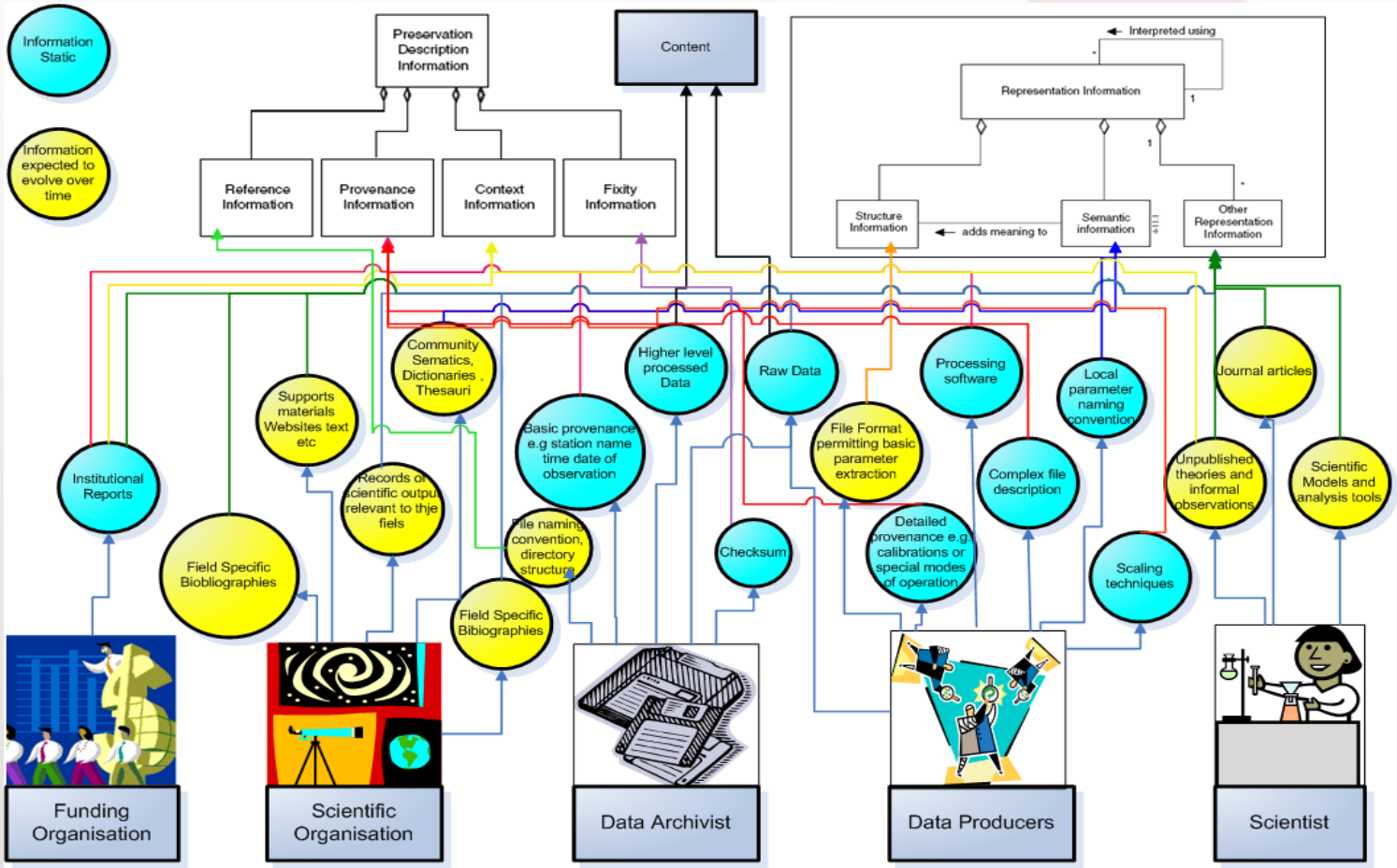


Scenario: Intelligibility-aware Packaging



- 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



Threat	Requirement for solution
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future	
Loss of ability to identify the location of data	
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	
The ones we trust to look after the digital holdings may let us down	

Accelerated Lifetime tests

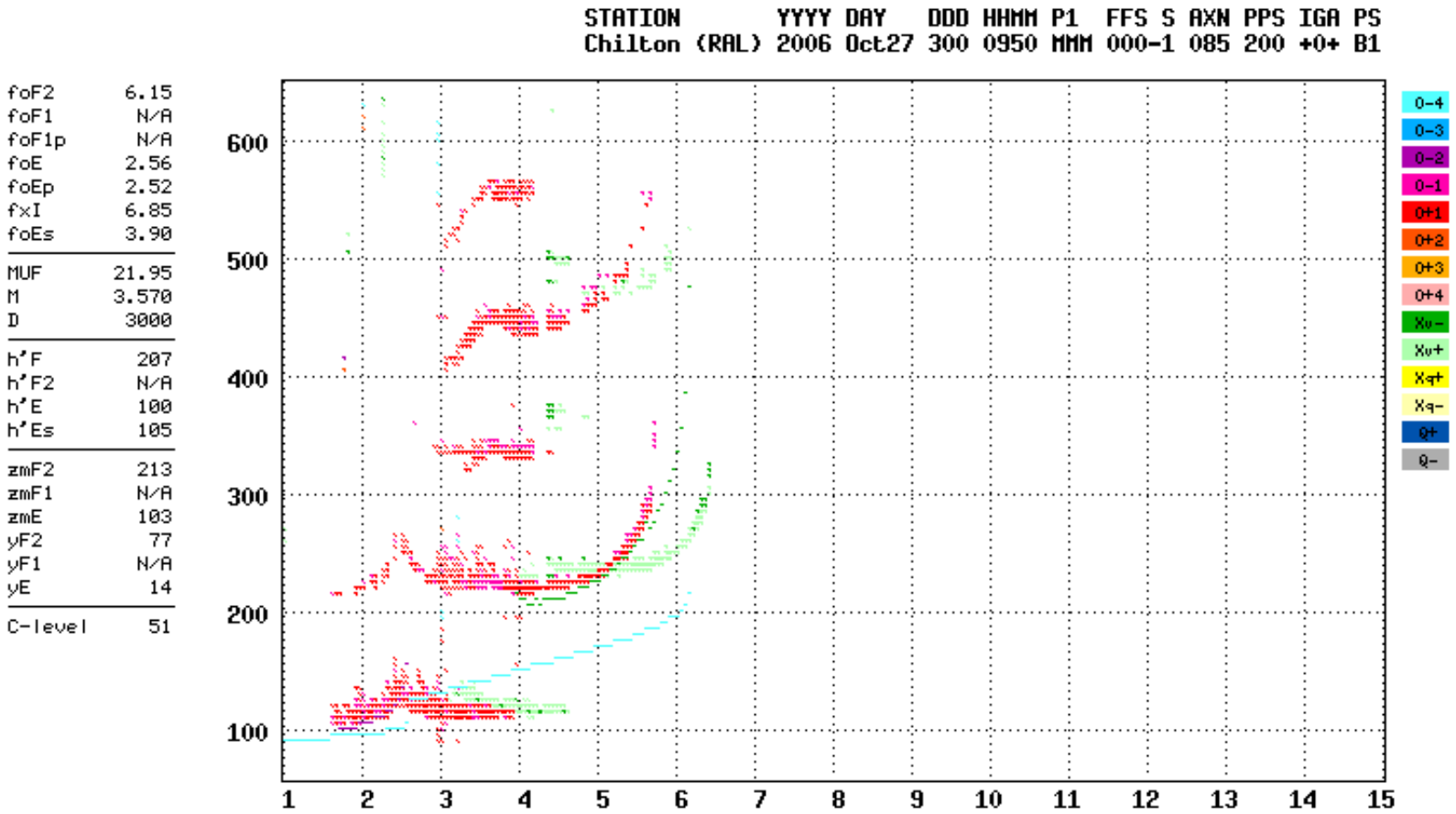
As part of the validation the CASPAR tested simulated the following:

- **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

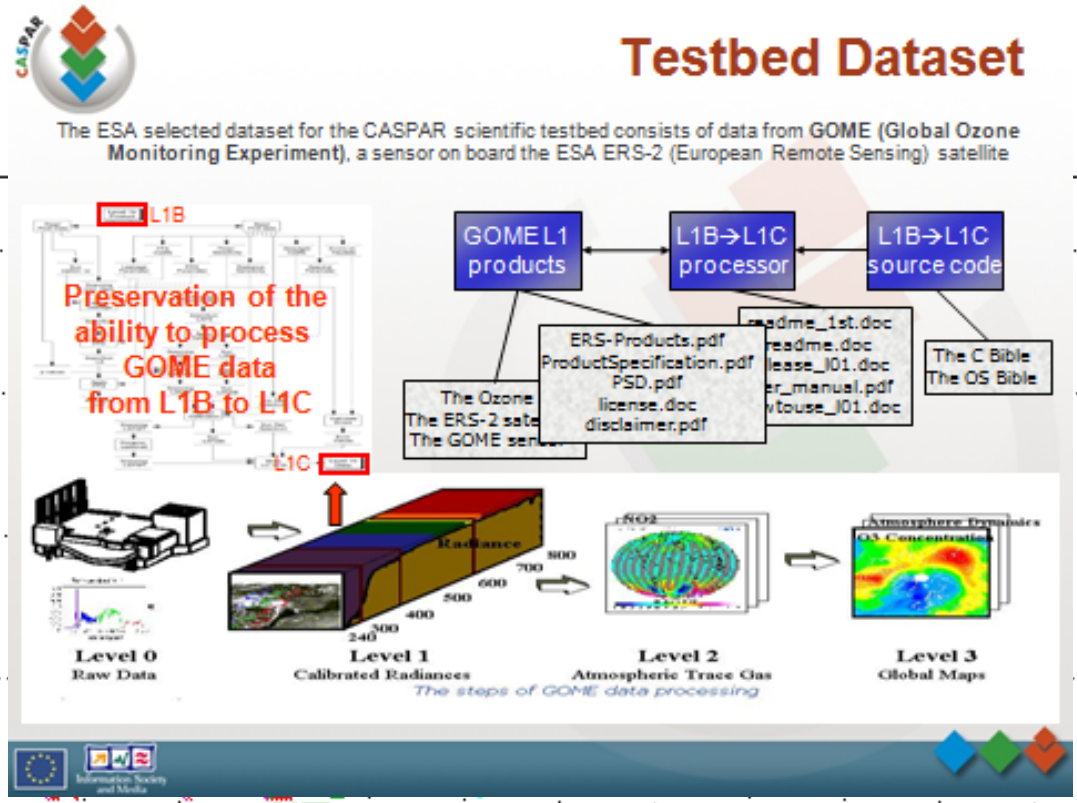
Threat	STFC	ESA	UNESCO	IRCAM	UnivLeeds	CIANT	INA
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	☑		☑	☑	☑	☑	
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	☑	☑	☑	☑		☑	
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	☑			☑			☑
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future							☑
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	☑						

STFC Testbed – various STP data



ESA testbed

foF2	6.15
foF1	N/A
foF1p	N/A
foE	2.56
foEp	2.52
fxI	6.85
foEs	3.90
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
yF2	77
yF1	N/A
yE	14
C-level	51

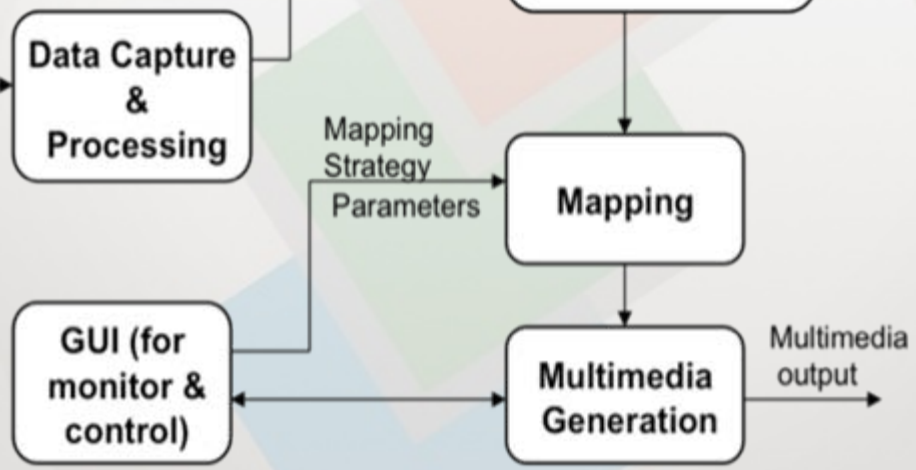


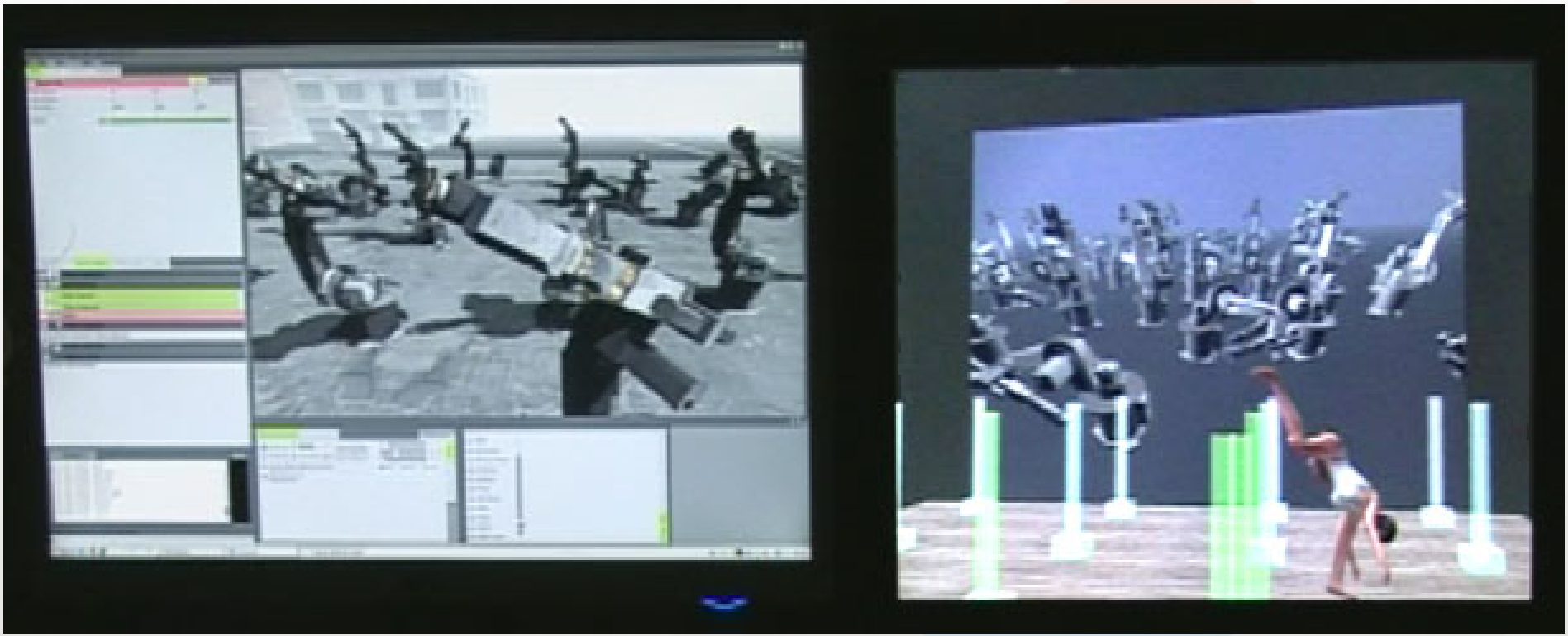
UNESCO testbed

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

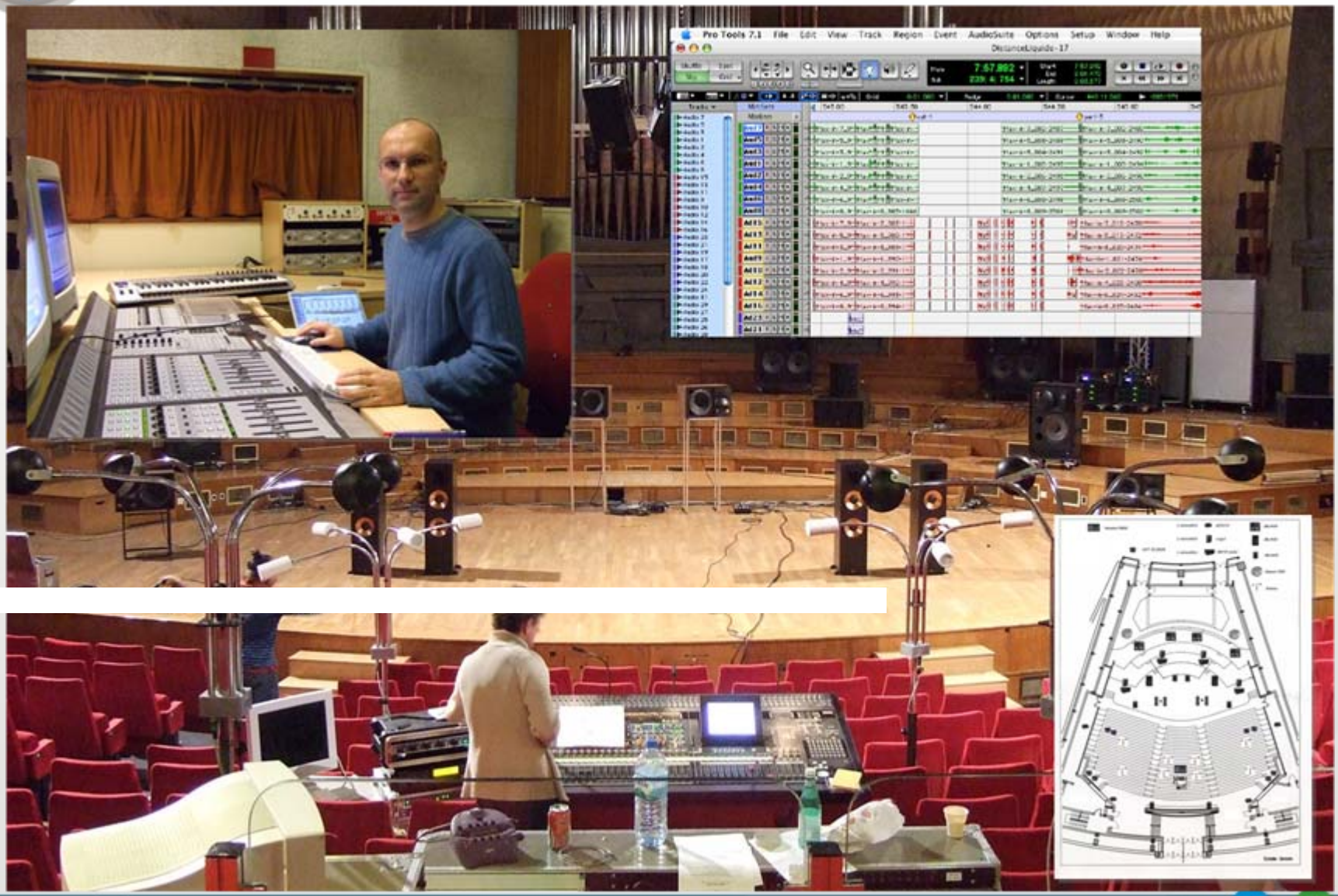
This is an elevation grid (height map) of the area where Villa Liva is located. It is an ASCII file in the ESRI GRID file format

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.


Pro Tools 7.1 File Edit View Track Region Event AudioSuite Options Setup Window Help

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)

Links

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
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Overview.aspx

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