CASPAR: Early results and future goals

David Giaretta

CASPAR aims

- **Produce tools and techniques to support digital preservation and make it easier to share the cost**
  - **must** be relatively easy to use
  - **must** have a low "buy-in" in terms of effort required for adoption
  - **must** avoid requiring wholesale change of everyone else's systems
  - **must** be decentralised and reproducible so that it can live on after the formal end of the **CASPAR** project
  - **must** be "**preservable**"
  - **must** be open: open source, open standards
- Cannot do everything but should do something broadly useful
- Working closely with the UK **Digital Curation Centre**

Digital Preservation…

- Easy to do…
- …as long as you can provide money forever
- Easy to test claims about tools…
- …as long as you live a long time

Validation

- Demonstrate theoretical basis
- Accelerated lifetime tests
  - Hardware
  - Software
  - Designated Community
- Increased trustworthiness
  - Measured using draft Certification Standard

Digital Preservation

- Need to preserve information & **knowledge** – not just “the bits”
  - Documents, videos are *rendered* – simple?
  - Data – must be processed - harder
- Need to manage knowledge to keep archives alive through time
  - Preservation is a process, not a one-time event
  - Preservation is expensive – costs need to be shared
    - The alternative is money – endless supplies of money
- Open Archival Information Systems Reference Model (ISO 14721) provides a general conceptual framework

Immediate benefits of Digital Preservation: Use of Unfamiliar Data

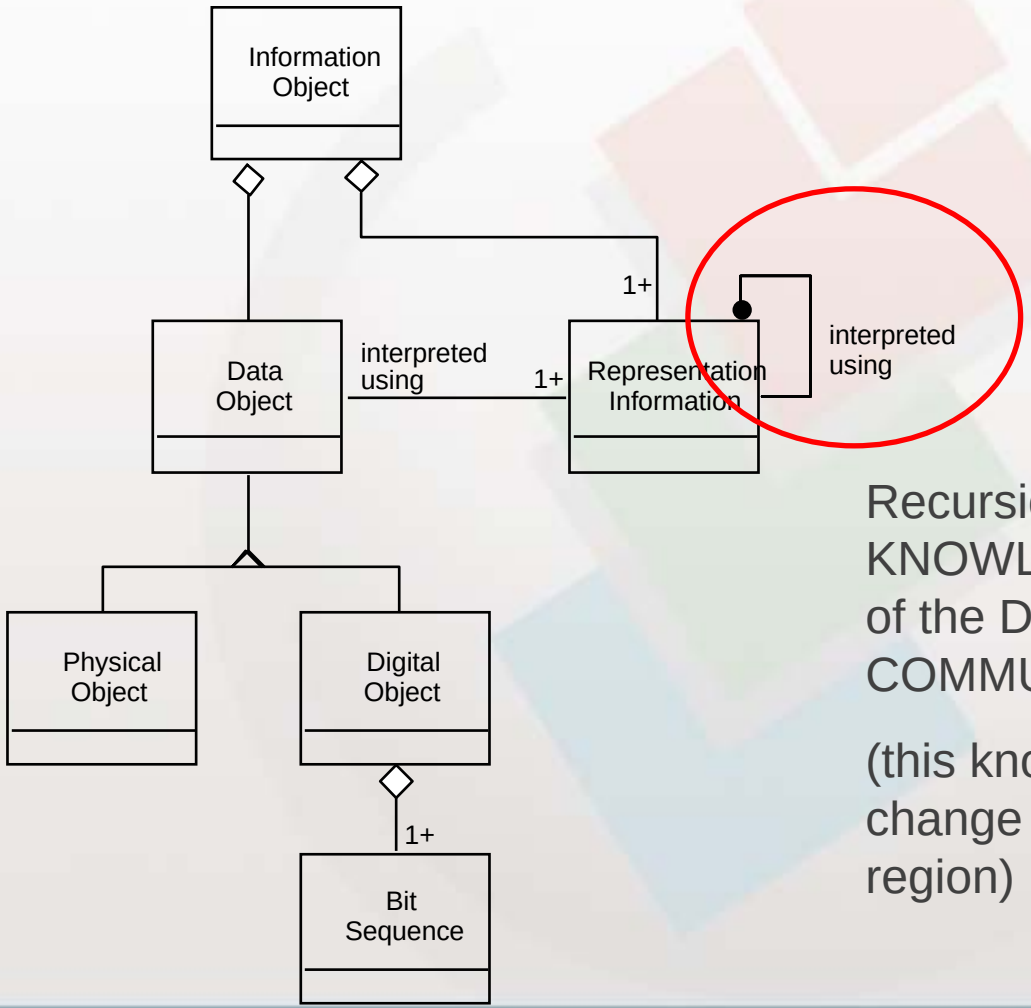
- Global Cyber-Infrastructures allow users to find and try to use data from many sources
  - Some sources will be familiar
  - **Most** available sources will be **unfamiliar**
- How can one be sure that the unfamiliar data is used correctly
- Garbage in – garbage out
- Need to be able to deal with unfamiliar data whether it is contemporary or old (preserved)

OAIS Reference Model

- **ISO 14721** : Reference Model for an Open Archival Information Systems (OAIS). *http://public.ccsds.org/publications/archive/650x0b1.pdf*
- An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.
- **Long Term Preservation**: The act of maintaining information, in a correct and Independently Understandable form, over the Long Term.
- **Long Term** is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community.
- **Designated Community:** An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.
- Has sufficient documentation to allow the information to be **understood and used** by the **Designated Community** without having to resort to special resources not widely available, including named individuals.

OAIS OAI

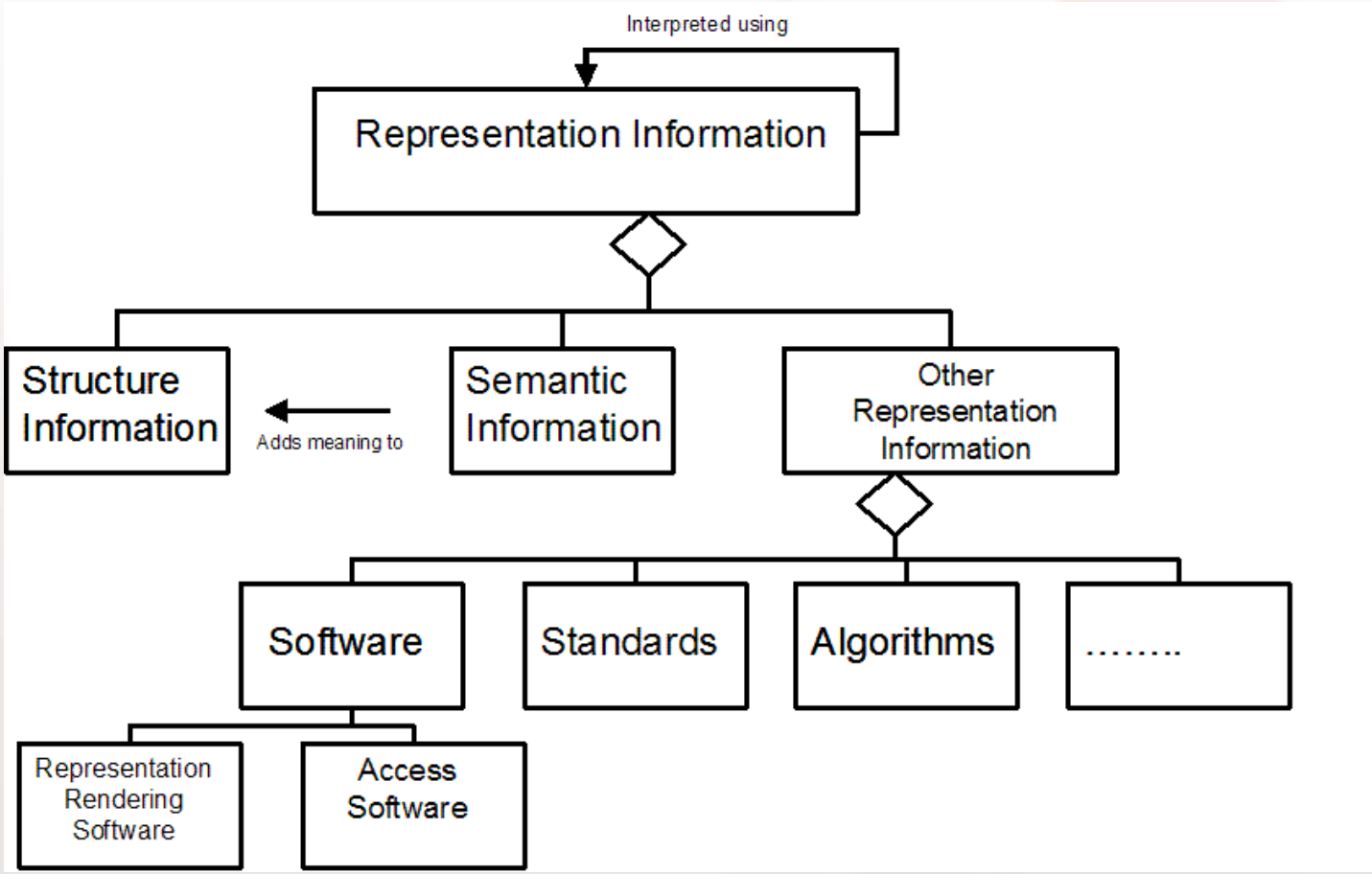
OAIS Information Model

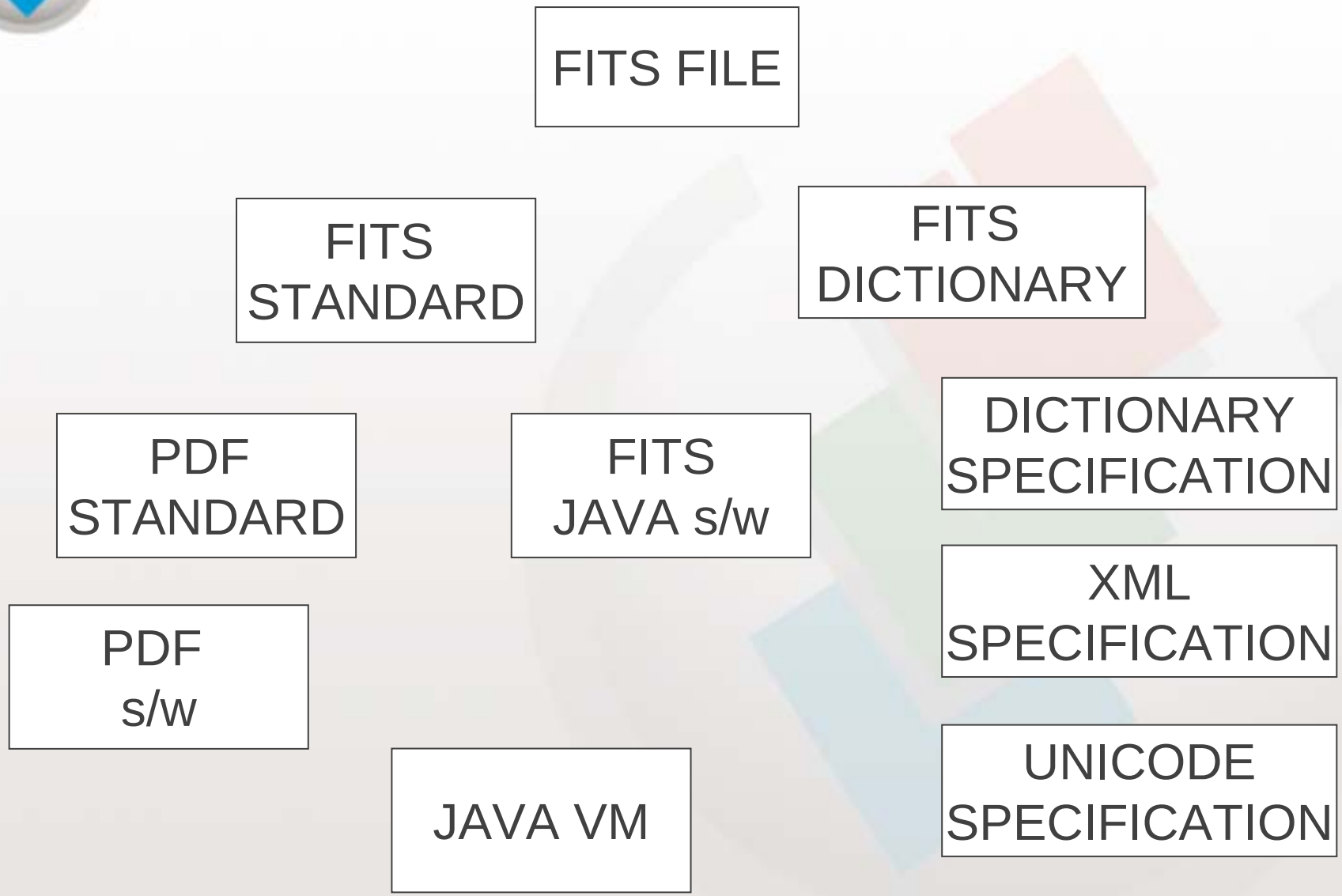


Recursion ends at KNOWLEDGEBASE of the DESIGNATED COMMUNITY

(this knowledge will change over time and region)

Rep.Info. Classification




FITS FILE
FITS STANDARD
FITS DICTIONARY
PDF STANDARD
FITS JAVA s/w
DICTIONARY SPECIFICATION
XML SPECIFICATION
PDF s/w
JAVA VM
UNICODE SPECIFICATION

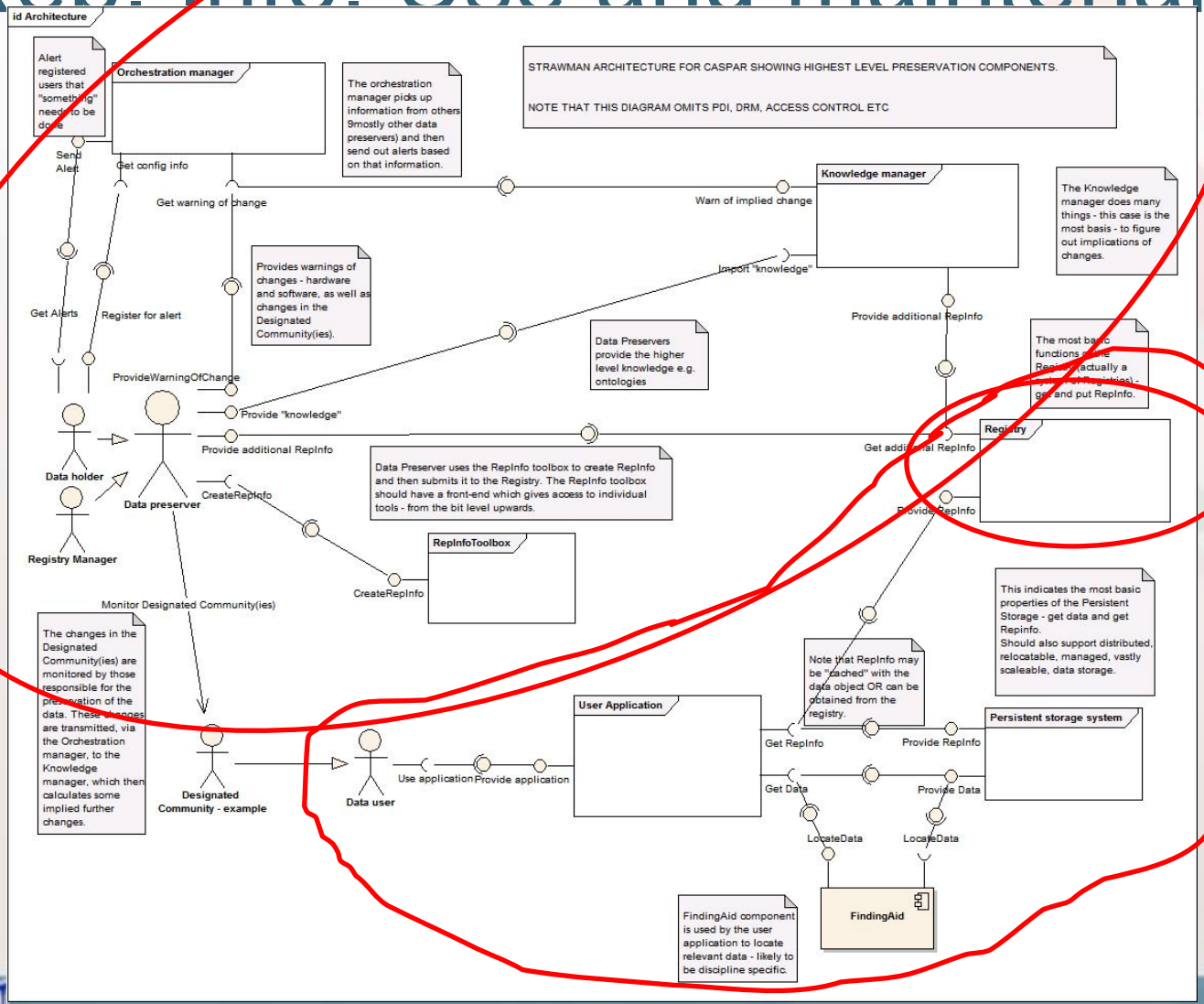
Representation Information

- The Data Object is “interpreted using” the Representation Information (RepInfo)
- The Reference Model is designed to ensure that an OAIS is **not** set the impossible task of having to provide **all** possible RepInfo **immediately**
- Hence:
  - Take account of the Designated Community and its associated Knowledge Base
- The amount of RepInfo is not fixed
  - Additional RepInfo will be needed over time

Early Results

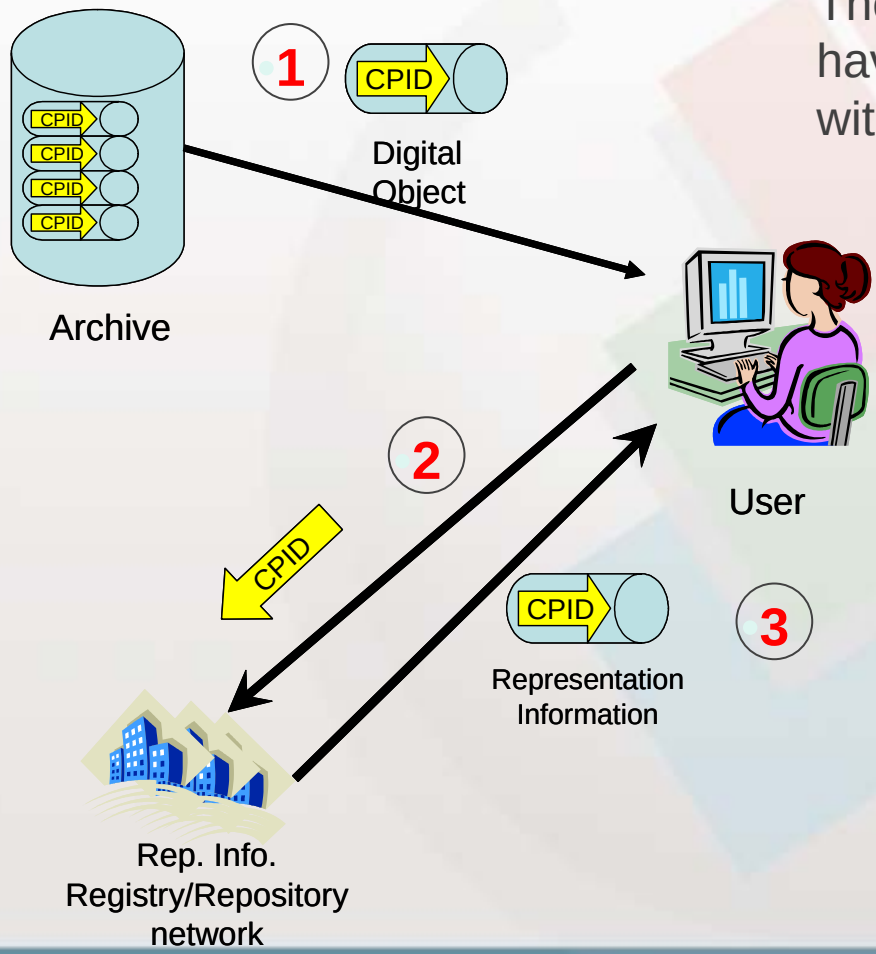
- High level architecture for sharing cost and access to Representation Information
- Detailed examinations of specific datasets to understand what is really needed to keep them understandable and usable

Rep. Info. Use and maintenance



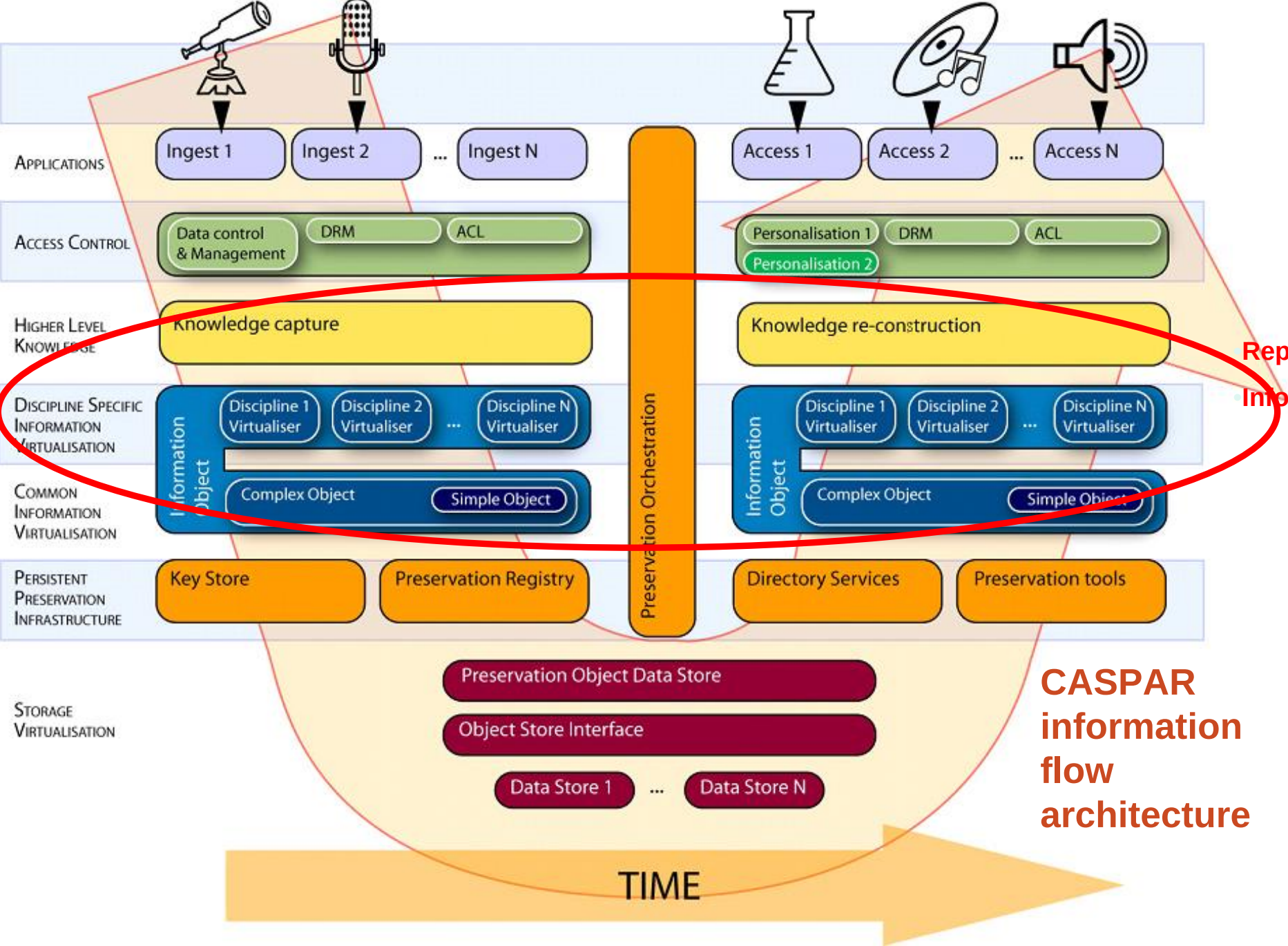
Registry for Representation Info

- **1** – User gets data from archive. Data has associated Curation Persistent Identifier (CPID)
- **2** – User unfamiliar with data so requests Rep.Info.using CPID
- **3** – User receives Rep.Info – which has its own CPID in case it is not immediately usable



The Digital Object could have RepInfo packed with it, as well as CPID

Support automated access & processing

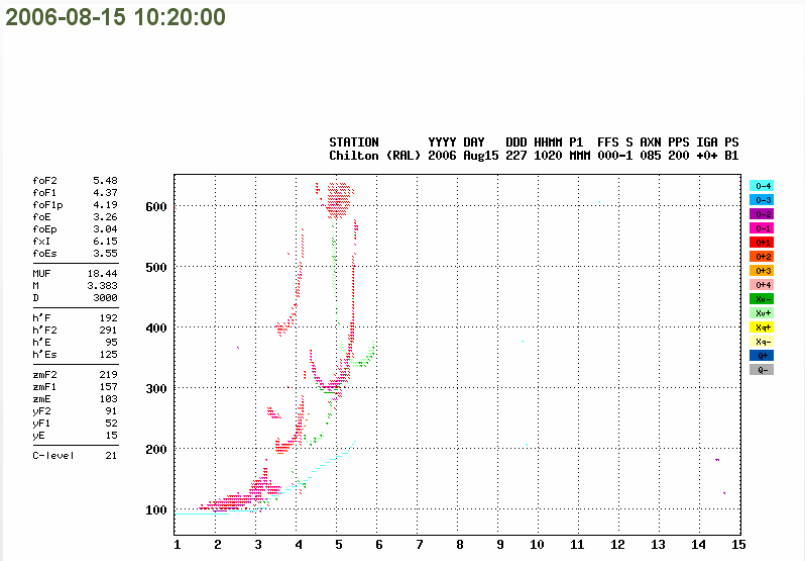


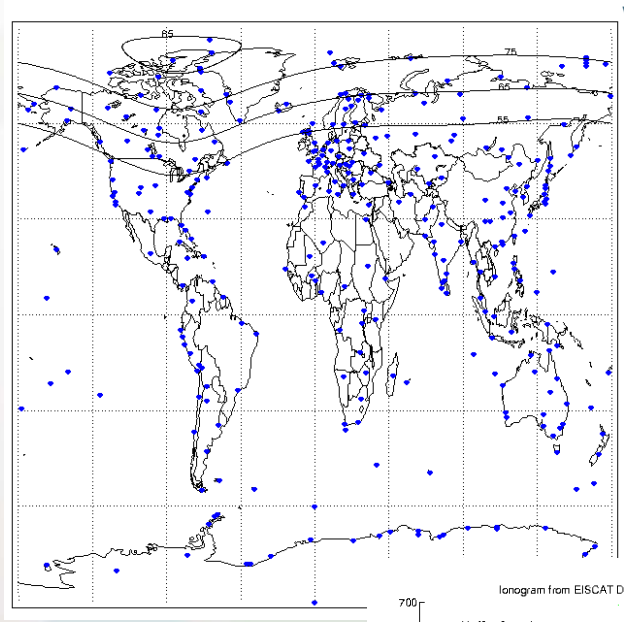
CASPAR information flow architecture

CASPAR Testbeds

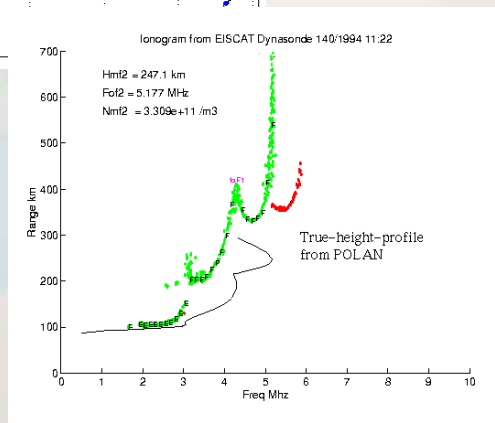
- Three testbeds
  - Cultural: UNESCO
  - Performing Arts: INA , IRCAM
  - Scientific: ESA and CCLRC
- Complex, multi-source, multifaceted data
- Many common preservation & evaluation & validation issues
- Some specific requirements on preservation (technical, delivery, legal)
  - Specific user communities/ Knowledge bases
- Also test the OAIS model

Science: CCLRC example





World map of ionosondes



Ionosonde data

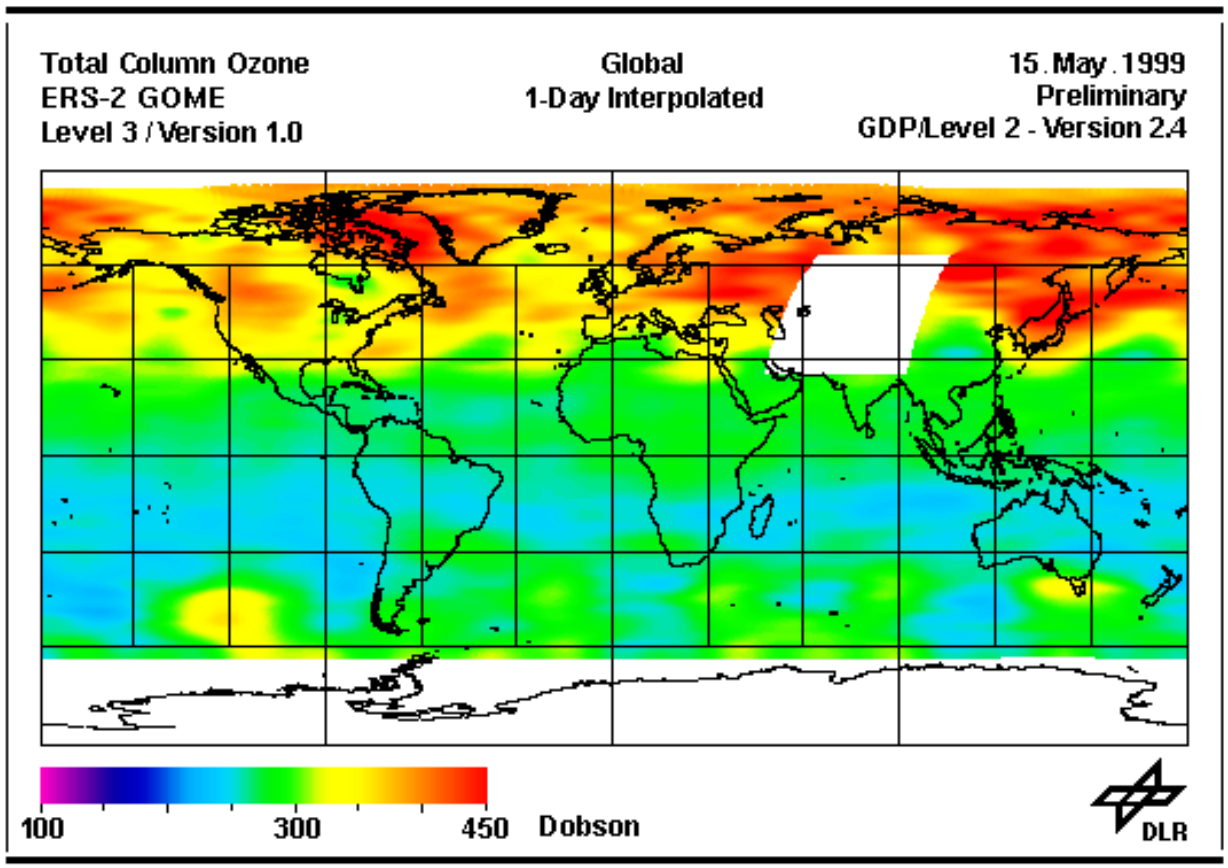
Some Issues

- Difficult to derive physical quantities from data
  - Can be analysed in multiple ways
  - Raises fundamental questions about Representation Information
- Common automated method is proprietary
  - Data structure also proprietary
  - Paper documentation - restricted access
- Provenance and trust

ESA example

GOME

Global Ozone Monitoring Instrument on ERS-2



GOME data processing

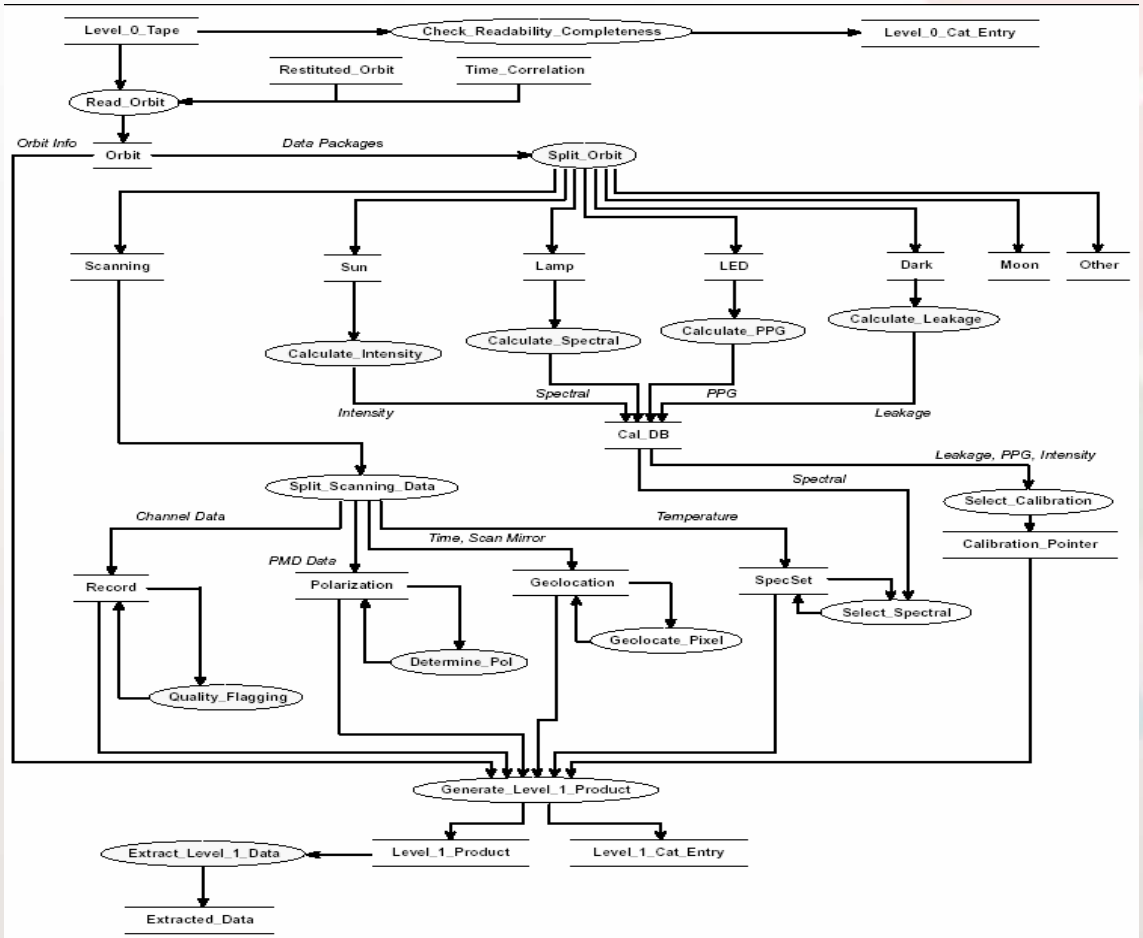
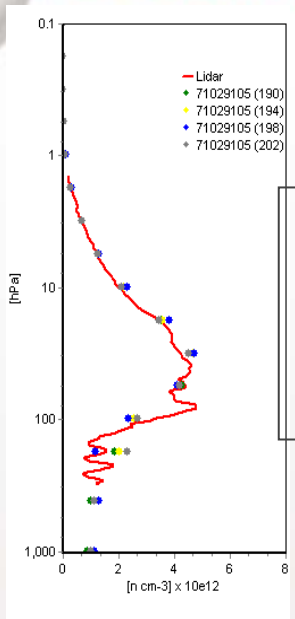
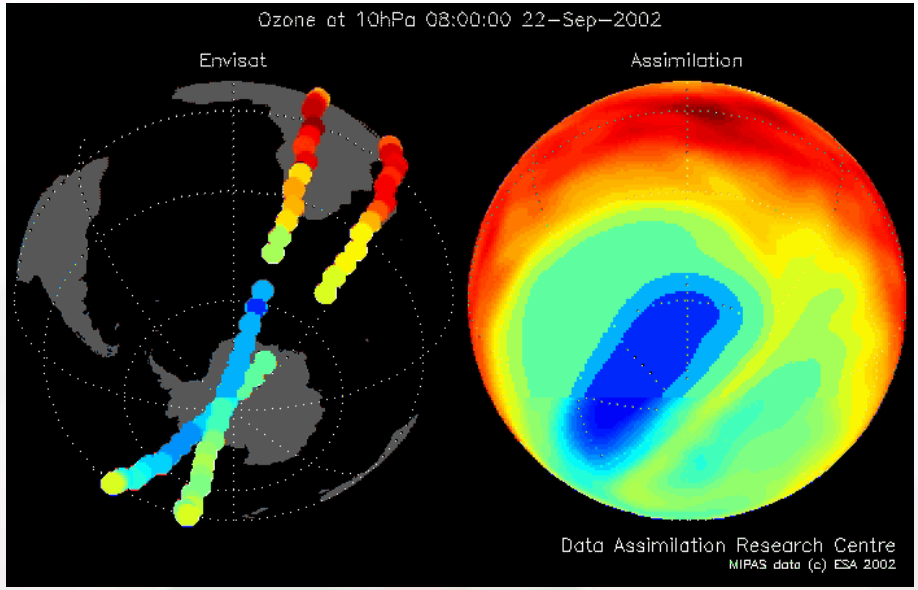
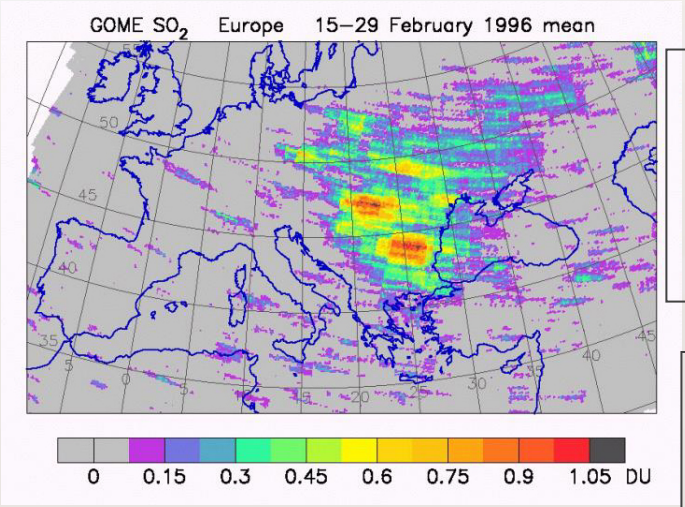


Figure 1: Functional Model Diagram for the Level 0 to 1 Processing

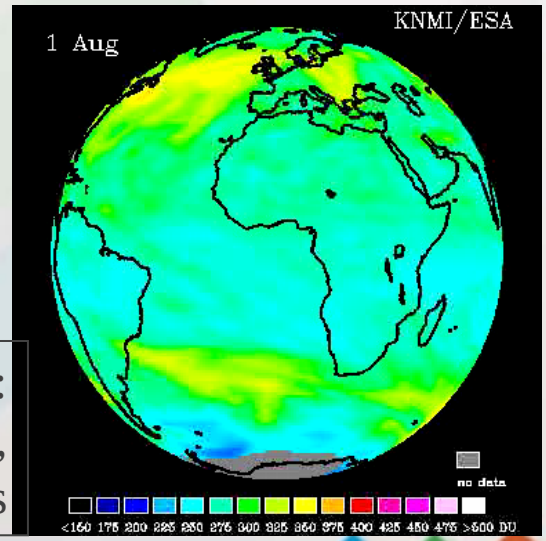


GOME Level 2 product:
Ozone profile at given location





GOME Level 3 product: Integration of time and space data



GOME Level 4 product: Integration of GOME, other data and models

Some Issues

- Provenance and Context of processed data

relationship to

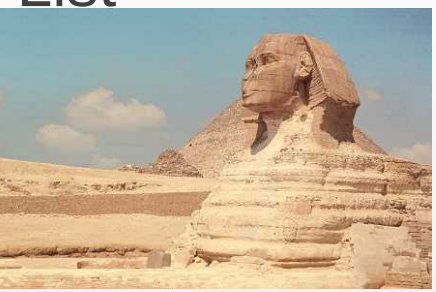
- Representation Information of raw data

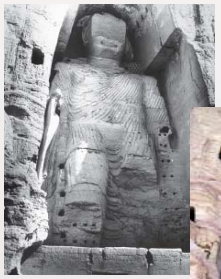
and

- Knowledge base of Designated Community

UNESCO examples

World Heritage List





Mandatory Documentation:

- 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)

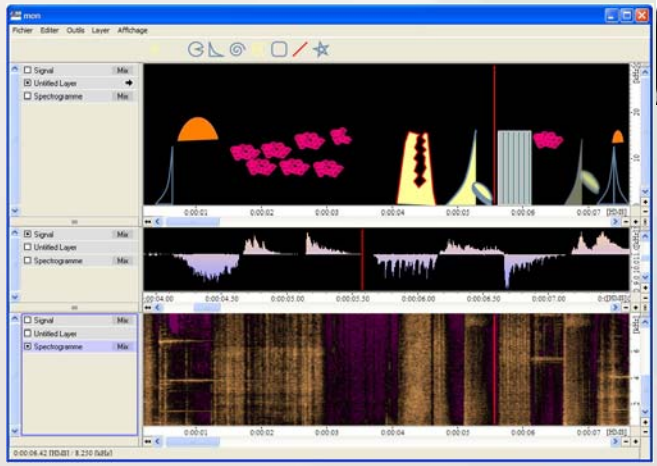
DATA:

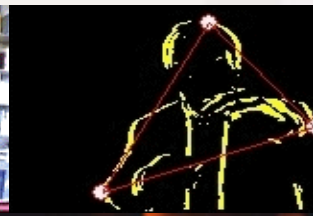
- Scanned documents and maps
- Aerial and close range photography (Digital photogrammetry)
- Monument measurements (Laser scanning)
- Satellite images (Remote sensing and image processing)
- Multi-scale digital cartography (Geographic information systems (GIS) and CAD)
- 3D models, virtual tours (Computer visualization)

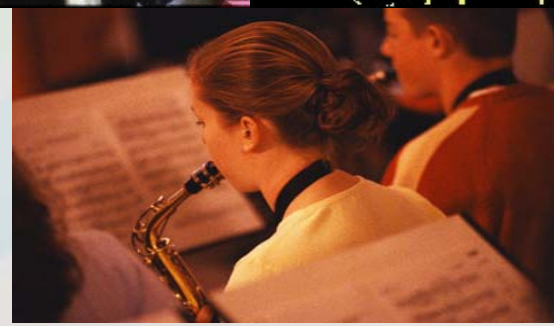
Performing Arts examples

Examples:

- Score
- MAX/MSP patches
- Additional instructions







Some Issues

- What is Preservation of “performability”?
  - Composer’s intention
- Authenticity
- Proprietary software and hardware
- Copyright
- Digital Rights Management

Shared Infrastructure

- Registries of Representation Information
- Persistent Identifier name resolvers
  - DOI? ARK? URL? – none are guaranteed
- Interfaces – support preservation and interoperability
- Standards – Preservation Description Information
  - Fixity, Provenance, Reference, Context
- Accreditation/Certification for repositories

Knowledge at the heart of preservation

- Knowledge driven approach
- Knowledge management to support long-term preservation of concepts/information including:
  - Single, complex, on demand, interactive objects
  - DRM
  - Authenticity
  - Access
  - Storage
  - Designated Community – descriptions
    - Knowledge base definition
    - ontologies

WHEN

- Component architecture and prototypes by month 12
- Framework architecture month 18
- Component integration months 24-30
- Testbed implementations months 30-36
- Project completion month 42

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

Home | CASPAR Project | Community | Publications | Events | Training | Press | Links | Join us

legend RSS

Welcome to the CASPAR Preservation User Community!

How can digitally encoded information still be understood and used in the future when the software, systems, and everyday knowledge will have changed? This is the challenge of CASPAR.

Digital information innervates modern civilization.

Yet digital information is extremely vulnerable. A huge amount of precious digital information created and stored all over the world becomes unaccessible every few years at a very fast pace. Think of losing official records, a museum archive, irreplaceable scientific data, or even a collection of family photos, and we realize digital preservation is affecting us all.

This website serves the project **CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval** - an Integrated Project co-financed by the European Union within the Sixth Framework Programme (Priority IST-2005-2.5.10, "Access to and preservation of cultural and scientific resources"), that started on 1 April 2006.

CASPAR will research, implement, and disseminate innovative solutions for digital preservation based on the **OAIS reference model (ISO:14721:2002)**.

The website provides official project documentation and material relevant to digital preservation and related disciplines. It also serves as an information and communication tool for the CASPAR Preservation User Community.

Community Members area

User ID cclrc

Password ******* login

Join the Community

To join the CASPAR Preservation User Community please use our Registration Form

Collaborate with CASPAR

Assist in the development of the CASPAR preservation components!

Contact us

To contact the CASPAR Project please use this form

Latest News

CASPAR Competing for Best IST Website Award!

Upcoming Events

www.casparpreserves.eu

Conclusions

- Science Data and Knowledge – needs more than just storing the “bits”
- Understanding and being able to process the vast amount of unfamiliar data which is available is hard
- It is expensive
  - Costs much be shared
- So far the Open Archival Information Systems Reference Model is OK
  - Many similarities can be exploited
  - Many subtleties need to be explored
- Watch this space