



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

CASPAR Overview

David Giaretta

International Conference on Digital Preservation at the
occasion of the retirement of J. Steenbakkers
Koninklijke Bibliotheek November 1-2 2007, The
Hague



Information Society
and Media





Drivers

- Fragility of digitally encoded information increasingly appreciated as a major concern
- This concern applies to almost every aspect of life
- Action and support needed at many levels
 - **Community**
 - **Political/Funding**
 - **Technical**





Search for commonalities

Things which tend to be very specific to each archive:

- Access methods
- Selection criteria
- Appraisal criteria

These will not be addressed here – probably not sufficient common ground

Focus on preservation





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





Preservation vs publication/access

- Needs of access:

- **Responsive**
- **Sophisticated search techniques**
- **Users often familiar with the material**

Transient tools and technologies with changing demands and implementations

- Needs of Preservation:

- **Ensure the information trapped in the bits is authentic and understandable**
 - To the Designated Community

Not transient





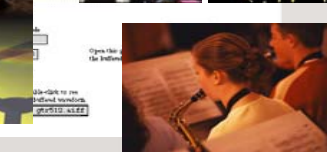
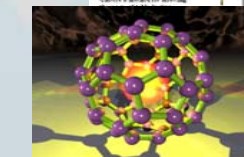
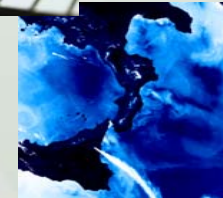
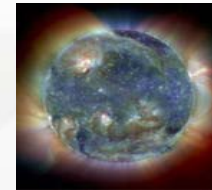
CASPAR Project

EU FP6 Integrated Project

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

Started April 2006, for 42 months

David Giaretta is Co-ordinator

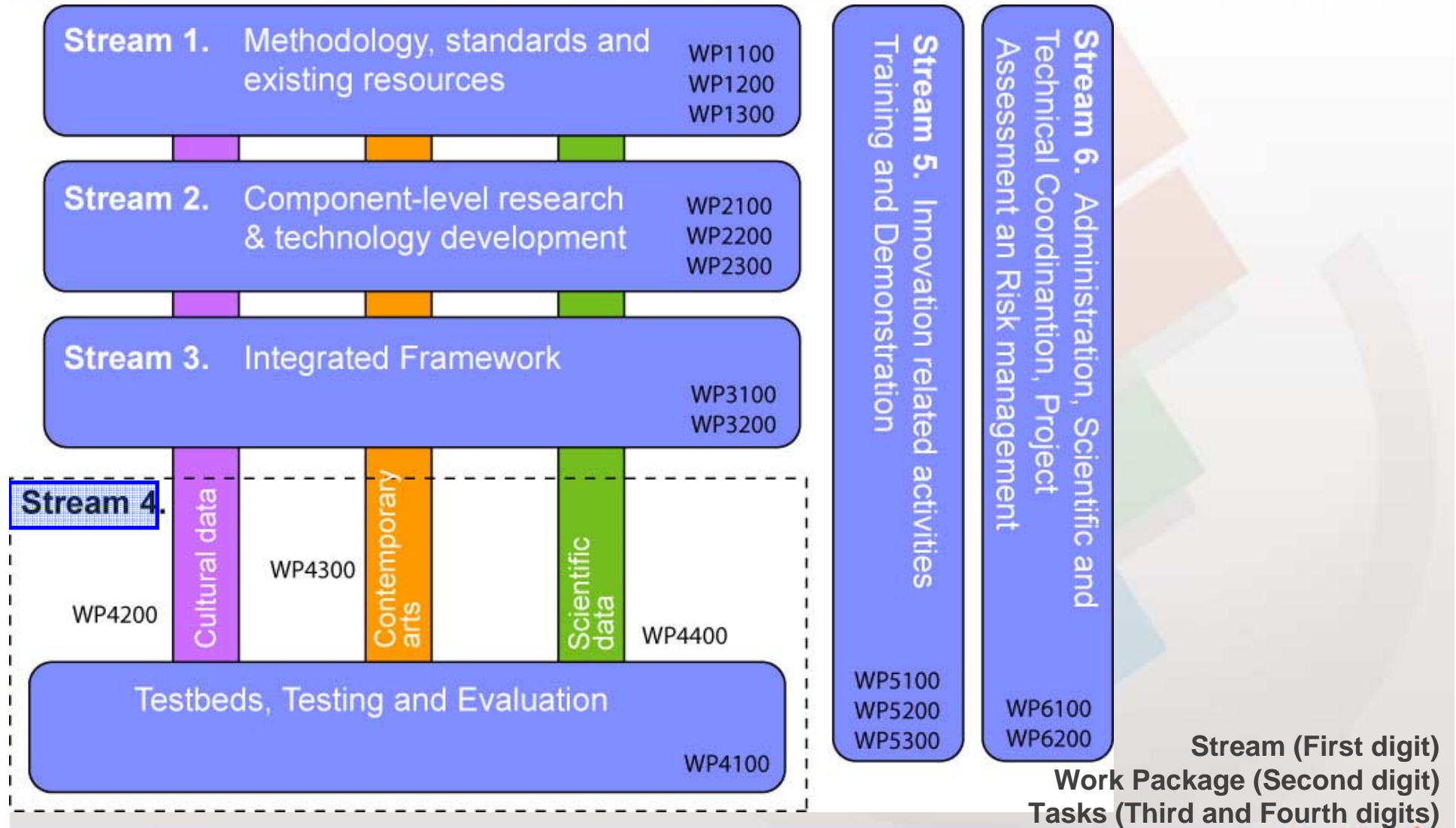


<http://www.casparpreserves.eu>





Project Structure

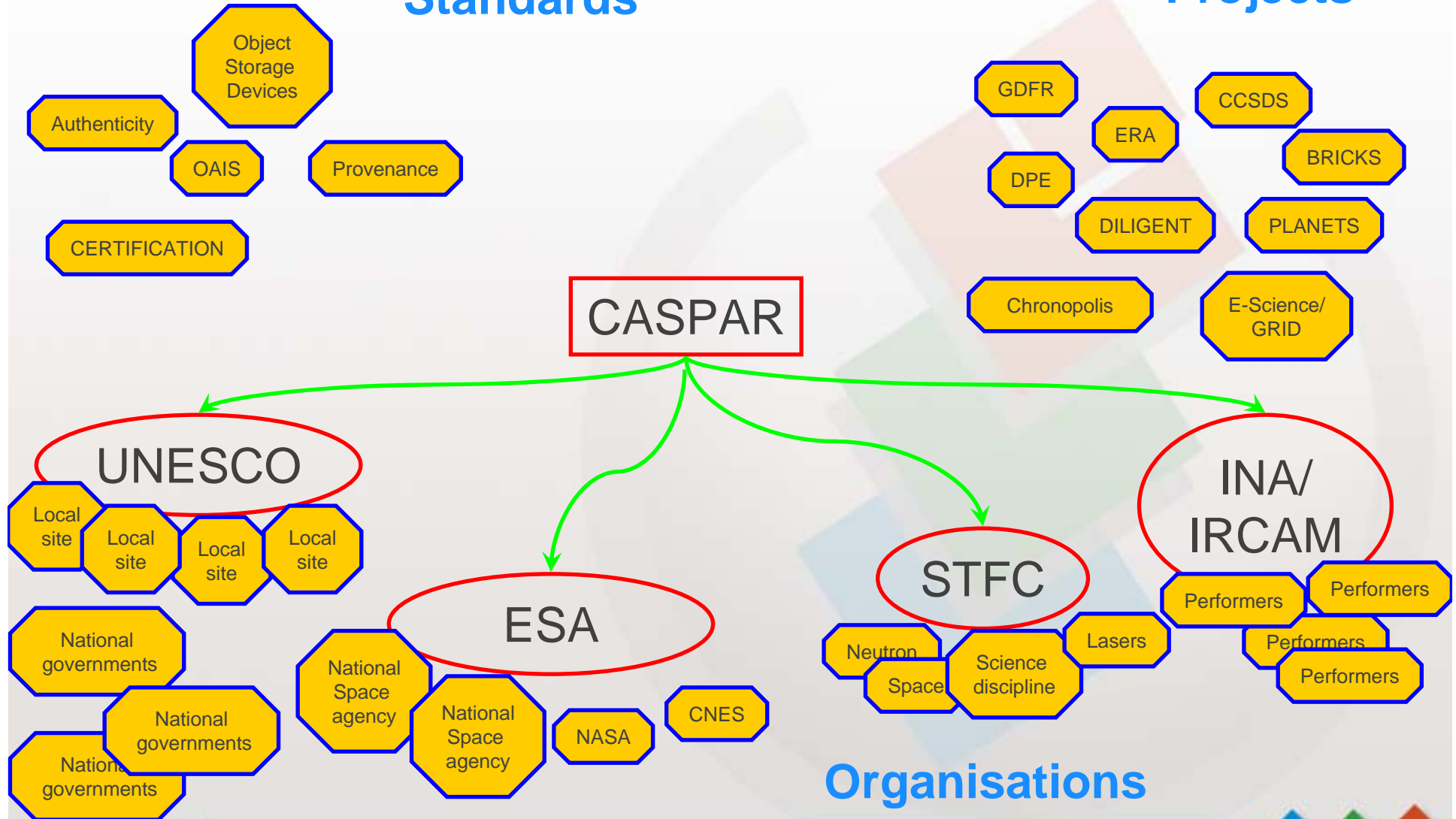




Some of the CASPAR links

Standards

Projects





CASPAR Aims

- Produce tools and techniques to support preservation of any and all digital objects 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**
- CASPAR cannot try to do everything
 - **Working closely with other projects**





Specific Objectives

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 to and use 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





Why CASPAR is special

- Digital preservation is hard
 - **OAIS view especially hard**
 - No organisation/project can guarantee its own longevity
 - Reduction of risk of losing information – cannot guarantee that nothing will be lost
 - **In the end it depends on money and interest**
 - Need to be able to share the load
 - Exploit OAIS concepts to the fullest extent
 - Broadly applicable
 - Need evidence of effectiveness
- Test with science, cultural heritage and performing arts data





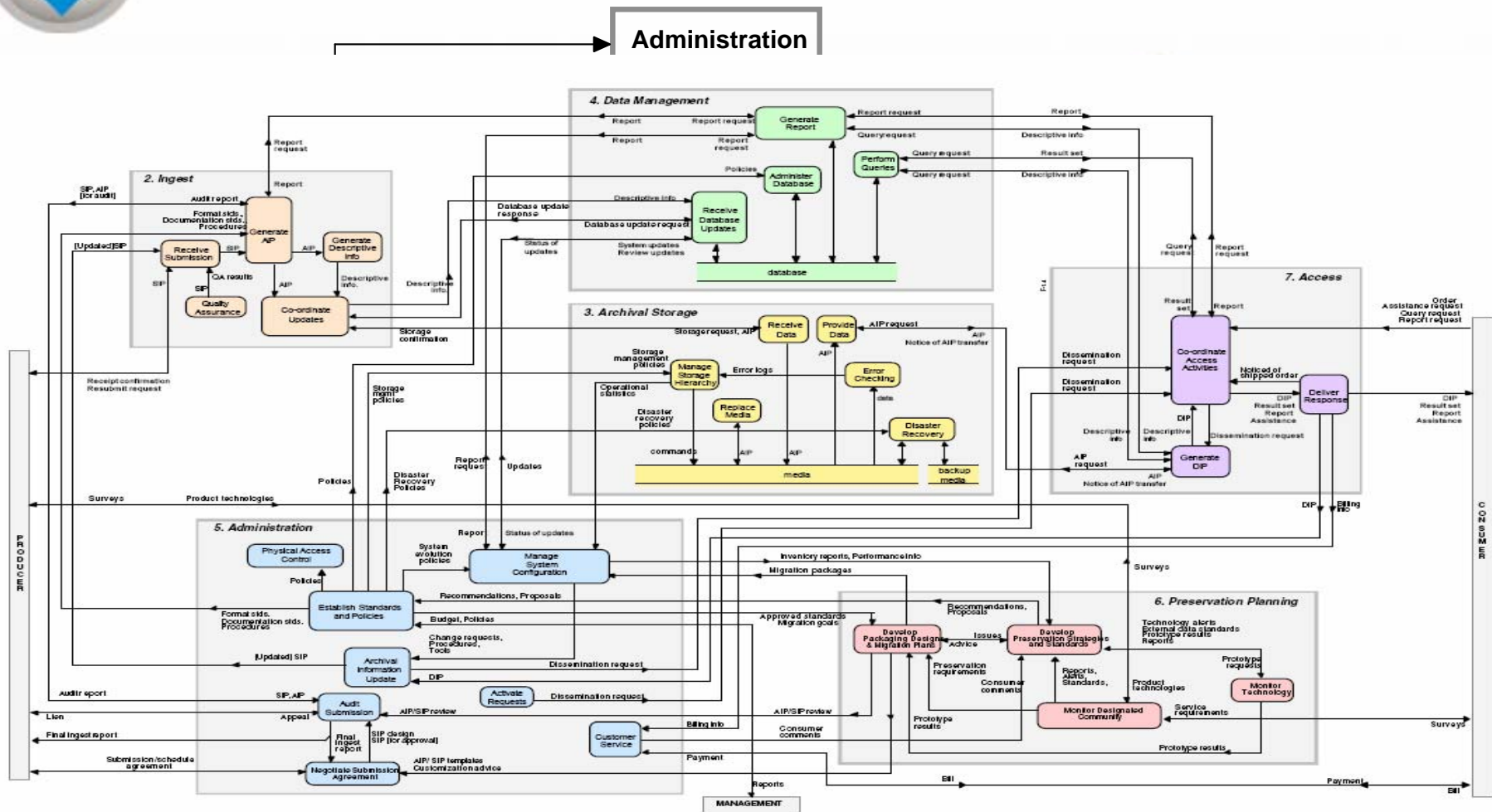
OAIS (ISO14721)

- Open Archival Information System Reference Model
 - **referenced in just about any serious work on digital preservation**
- 5 year ISO review underway
 - **minor corrections and updates**
 - **No major changes**
- Revised version due early 2008





OAIS Functional Entities



SIP = Submission Information Package
 AIP = Archival Information Package
 DIP = Dissemination Information Package



Key OAIS (ISO 14721) Concepts

- Claiming “This is being preserved” is untestable
 - **Essentially meaningless**
- How can we make it testable?
 - **Claim to be able to continue to “do something” with it**
 - Understand/use
 - **Need Representation Information**
- Still meaningless...
 - **Things are too interrelated**
 - Representation Information potentially unlimited
 - **Designated Community**
- Plus many other concepts to aid clarity



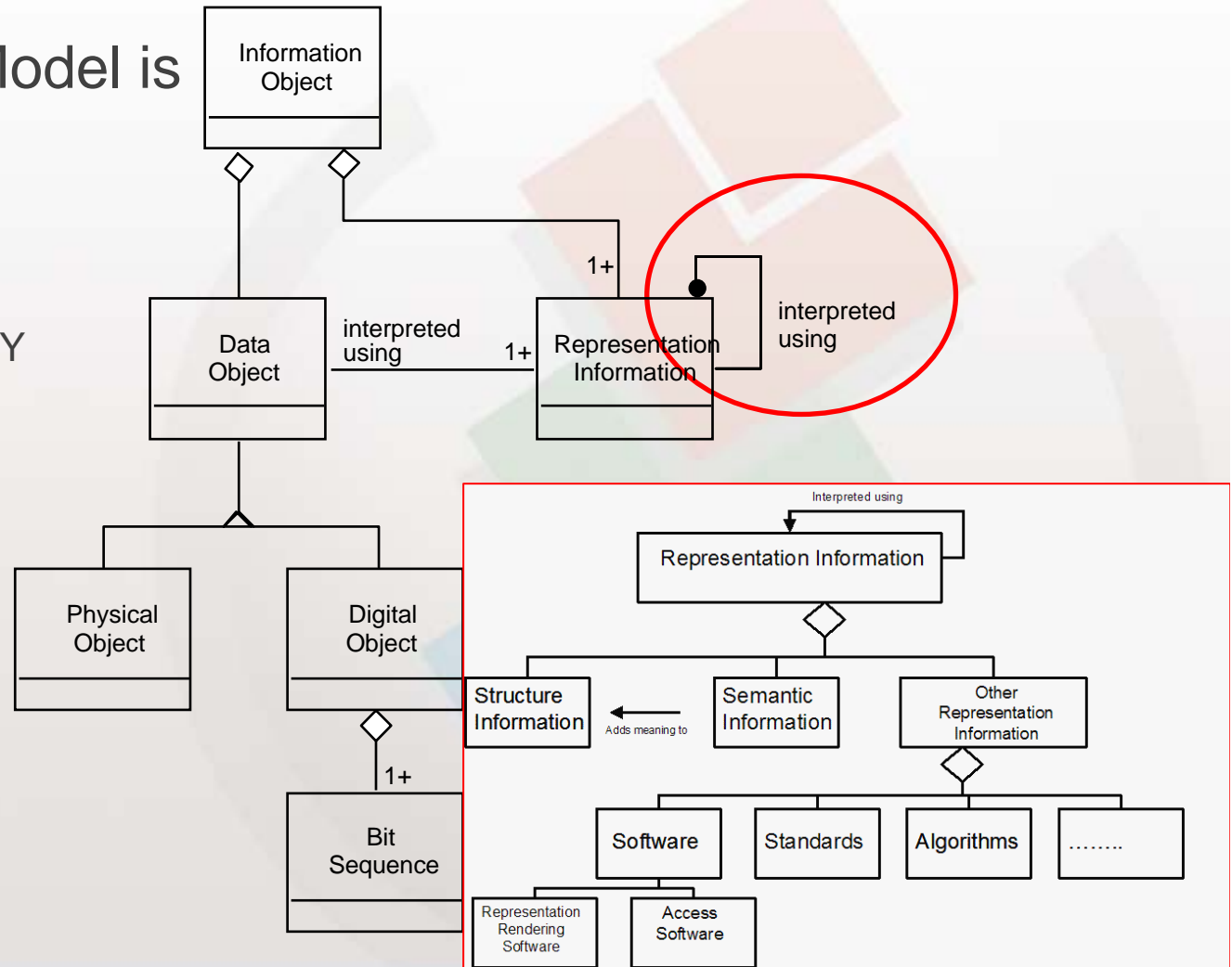


Information Model & Representation Information

The Information Model is key

Recursion ends at KNOWLEDGEBASE of the DESIGNATED COMMUNITY

(this knowledge will change over time and region)





Documents vs Data

- Need to preserve information & knowledge – not just “the bits”
 - Documents, videos are *rendered* – simple?
 - Data – must be processed – in new ways - harder
 - more Representation Information

Publication of data as well as documents





Data – OAIS view

A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

- In 2006, the amount of digital information created, captured, and replicated worldwide was 161 exabytes(161 billion gigabytes) -roughly 3 million times the information in all the books ever written!*
- Between 2006 and 2010, the information added annually to the digital universe will increase more than six fold from 161 exabytes to 988 exabytes*

IDC (2007) The Expanding Digital Universe





Just Format?

representation information rules

You have a file

JHOVE tells you it is WORD version 7

Format – necessary but not sufficient:

formats can be used for multiple purposes e.g. audio files
used to store configuration parameters





Rendered objects people

- Documents, articles, journals...
- Images
- Audio
- Video

A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

Semantics tends to be ignored





Possible mapping of terminology

“Rendered Object” terminology	OAIS terminology
FORMAT	the type of Representation Information which describes Structure (as distinct from Semantics)
METADATA	Package Description (or the Descriptive Information which is derived from it) – which is used in Access Aids including Finding Aids . Note that normally the term “metadata” covers much more but looking at usage in this area it seems clear that this more restricted meaning is understood.





Information is the important thing

- What information?
 - Documents.....
 - Data.....
- Original bits?
- Look and feel?
- Behaviour?
- Performance?
- Explicit/ Implicit/ Tacit

Information:

Any type of knowledge that can be exchanged. In an exchange, it is represented by data.

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. Long Term may extend indefinitely.

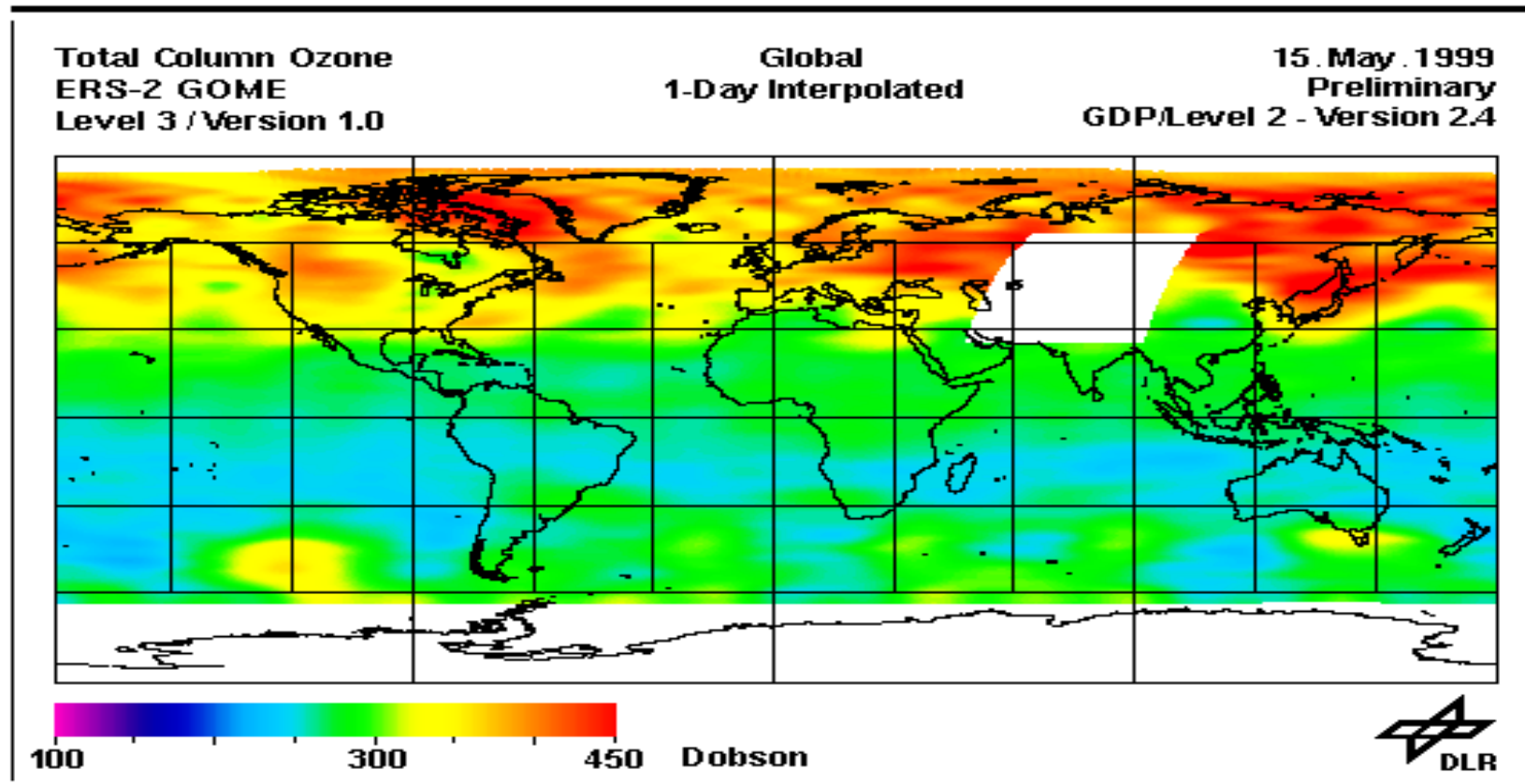
Ensure that the information to be preserved is Independently Understandable to (and usable by) the Designated Community.





Data...

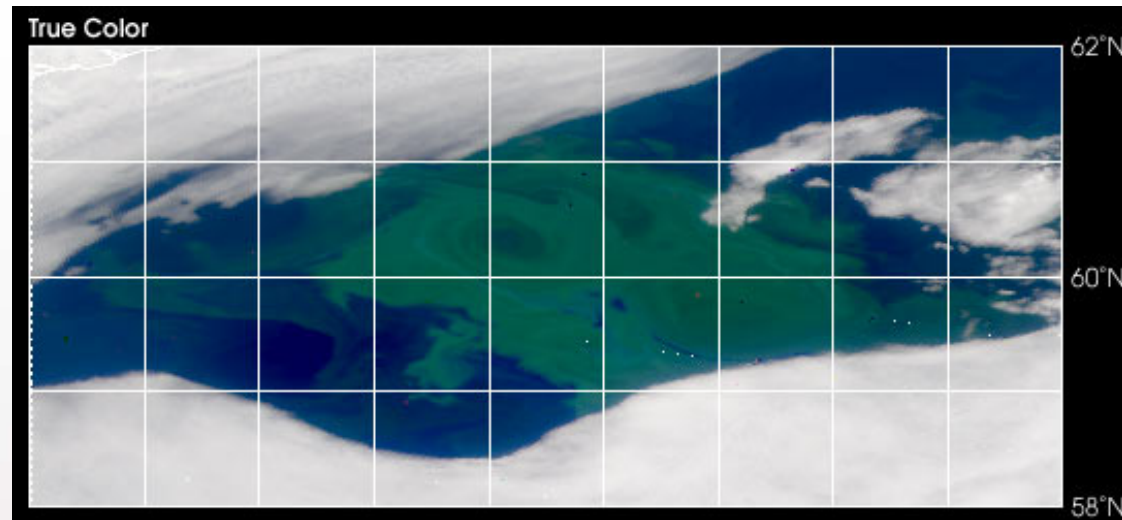
Level 2 GOME Satellite instrument data



Information Society
and Media



SEAWIFS IMAGE



Information Society
and Media



FITS FILE

FITS
STANDARD

FITS
DICTIONARY

PDF
STANDARD

FITS
JAVA s/w

DICTIONARY
SPECIFICATION

PDF
s/w

XML
SPECIFICATION

JAVA VM

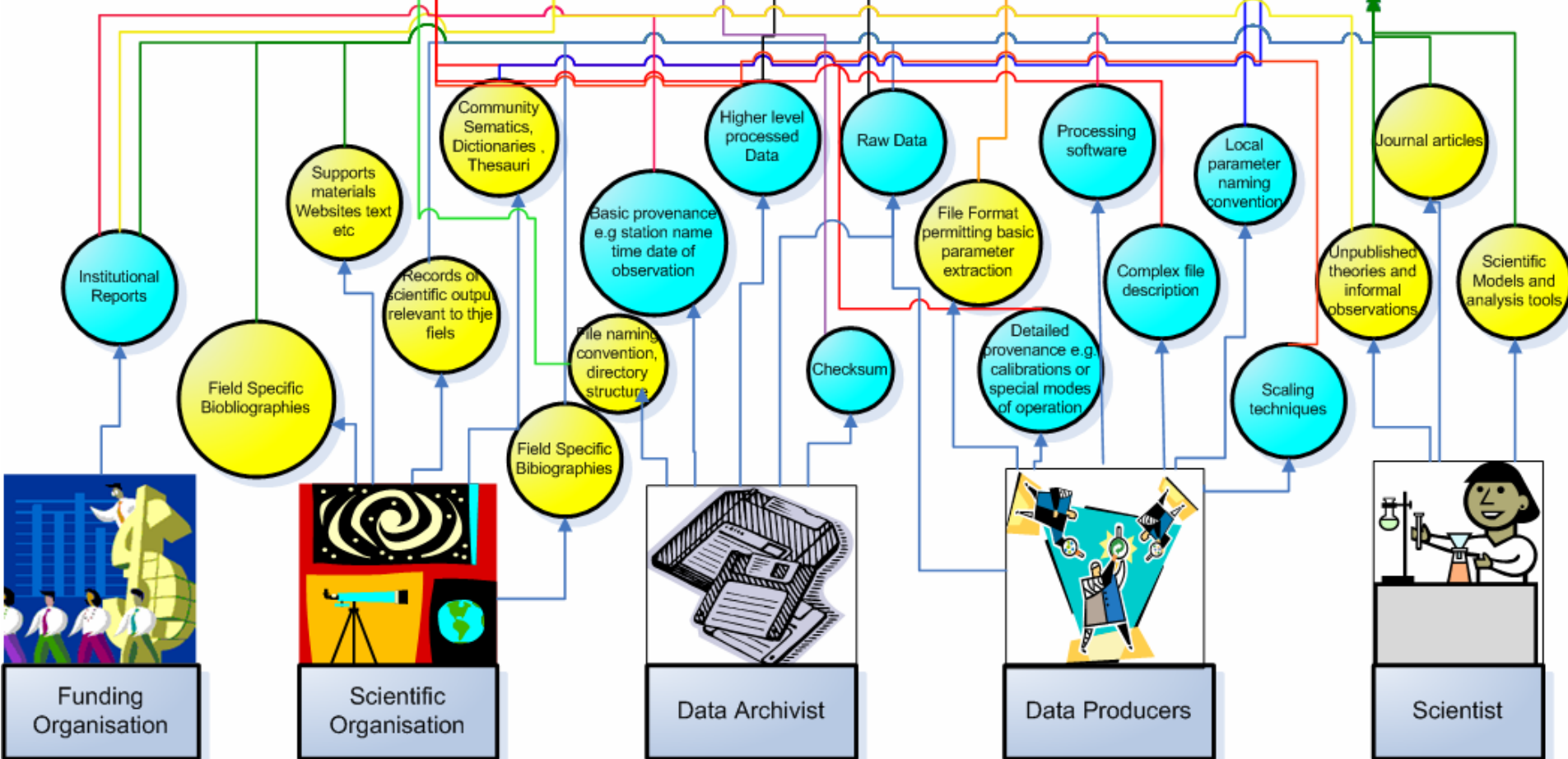
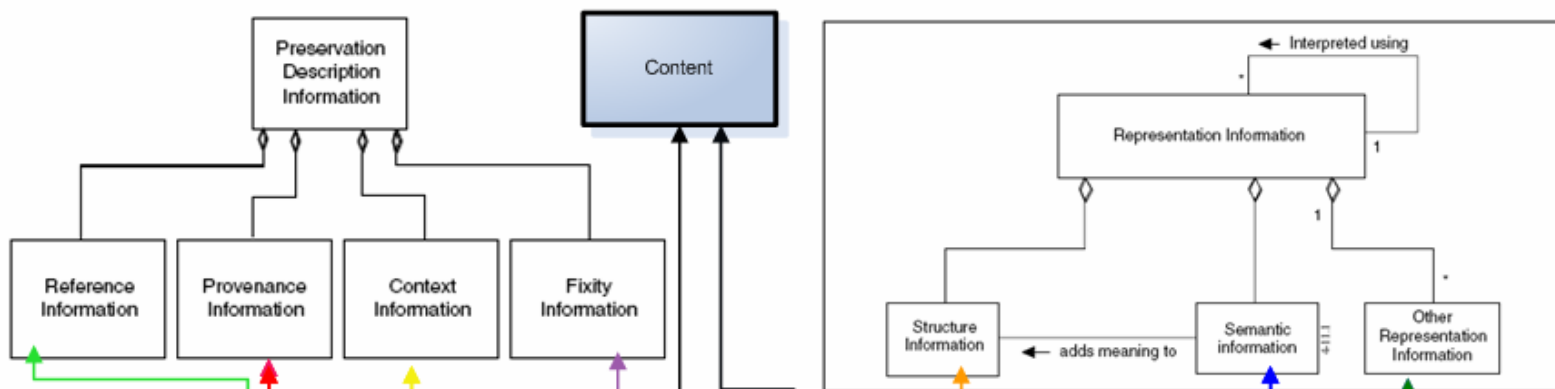
UNICODE
SPECIFICATION





Information Static

Information expected to evolve over time





“Levels” of Representation Information

- Checklist – not just “metadata” e.g.
 - **Representation Information (ReplInfo)**
 - **Preservation Description Information**
 - **Packaging Information**
- Exhaustive text documentation
- ReplInfo supporting:
 - **Interoperability**
 - **Automation**

Double payback

- **Applicable in “GRID” context**
 - usability now as well as later





Things change/disappear

- Software
- Hardware
- Environment
 - **E.g. Network links to related information**
- People
 - **What is “common knowledge”**

How can we ensure that the information trapped in the “bits” remains understandable despite all these changes?

How can a digital curator even be aware of these changes?





Validation

- How can we judge any proposed solution?

Live a long time

- CASPAR validation metrics:

- **Theoretic underpinning**

- **Testbed scenarios addressing real issues**

- No “hand-waving” – use what is there now

- Accelerated lifetime tests

- **Hardware and Software**

- **Environment**

- **People**

Evidence - not proof

- **Improved “trustability”/”certifiability”**





Technical Deliverables

Deliverables in place ready for implementation:

- D4101 – User Requirements and Testbed Scenarios
 - **Supported by detailed OAIS-based questionnaires**
- D1101 – State of the Art
 - **Follows OAIS-guided questionnaire**
- D1201 – Conceptual Model
- D1301 – Architecture
 - **UML diagrams supported by common UML tool**





Summary

- 17 member consortium
- Covering
 - **National and International Science, Performing Arts and Cultural Heritage data holders**
 - **Software expertise**
 - **Academic excellence**
- Focus
 - **OAIS**
 - **Preservation framework to share the preservation burden for all types of digital objects**
 - **Validation**





Links

- **CASPAR project**
 - **EU project on digital preservation – Science, Culture and Arts data**
 - Infrastructure, tools and detailed case studies – what does one need to actually “understand” the data?
 - <http://www.casparpreserves.eu>
- **Digital Curation Centre**
 - <http://www.dcc.ac.uk>
- **Audit and Certification group Wiki:**
 - <http://wiki.digitalrepositoryauditandcertification.org>
- **Alliance Permanent Access**
 - <http://www.alliancepermanentaccess.eu>

