



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

Trusted Data Repositories

David Giaretta

STFC

and

Director of CASPAR

and

Associate Director UK Digital Curation Centre



ADASS 24-26 Sept 2007



The need for Trustable Repositories

- Task Force on Archiving of Digital Information (1996) declared,
 - “a critical component of digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections.”
 - “a process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.”
- A recurring request in many subsequent studies and workshops



Digital Preservation...

- Easy to do...
- ...as long as you can provide money forever
- Easy to test claims about repositories...
- ...as long as you live a long time
- Reference Model for Open Archival Information System (OAIS) provides an approach





Key OAIS 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
- Many other concepts identified
 - Checklist – not just blanket term of “metadata”



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.

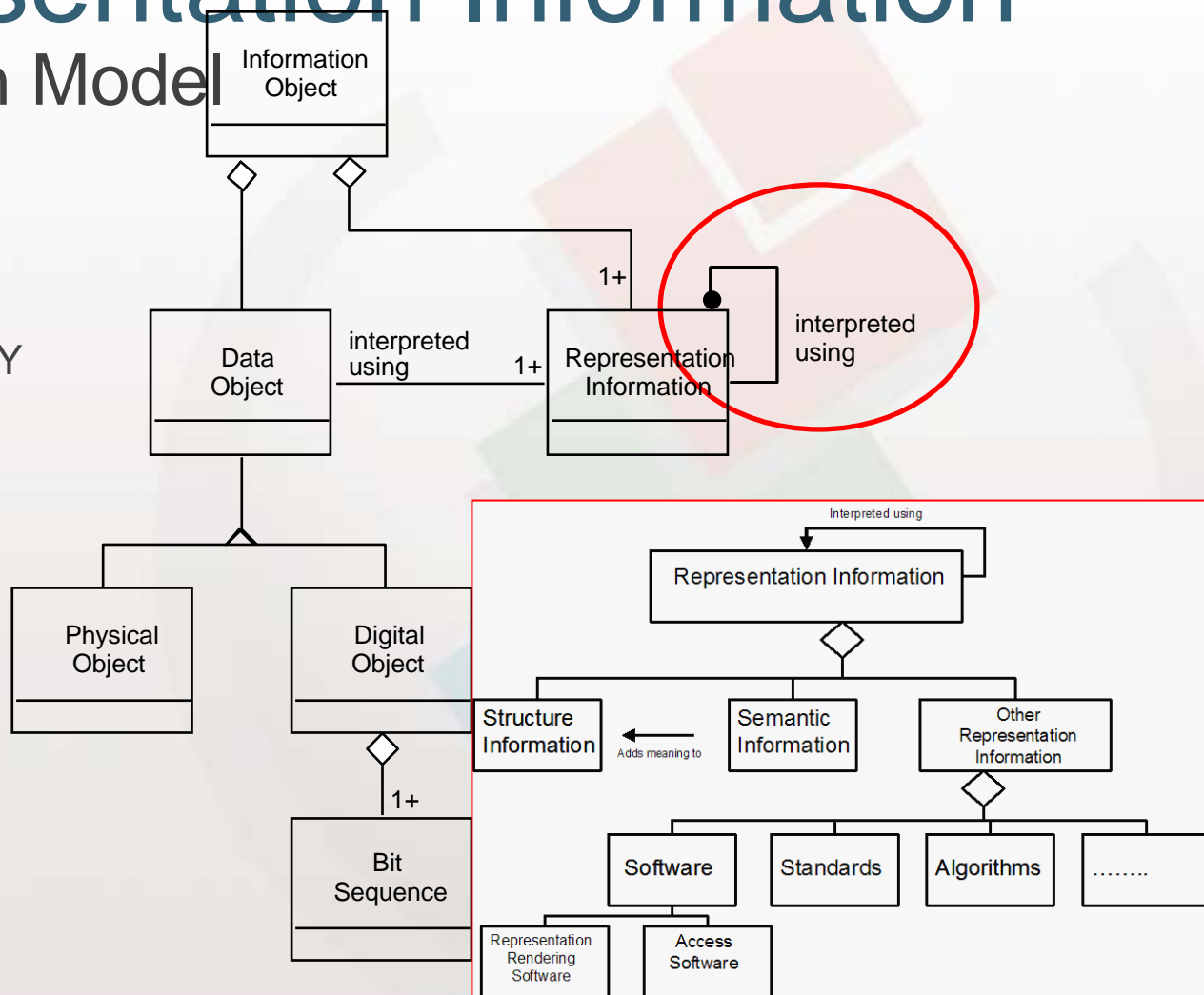
Ensure that the information to be preserved is Independently Understandable to (and usable by) the Designated Community.

Representation Information

The Information Model
is key

Recursion ends at
KNOWLEDGEBASE of the
DESIGNATED COMMUNITY

(this knowledge will change
over time and region)



FITS FILE

FITS
STANDARD

FITS
DICTIONARY

PDF
STANDARD

FITS
JAVA s/w

DICTIONARY
SPECIFICATION

PDF
s/w

XML
SPECIFICATION

JAVA VM

UNICODE
SPECIFICATION



Changes in hardware and software

Change of

- hardware
- operating system
- compilers/libraries/drivers

Affecting ability to run

- the Data Management System
- the specialised data processors
- the format/auxiliary data converters.



Changes in environment

- Change of software licence/copyright affecting ability to read data
- Evolution and merging of organisations affecting access rights
- Dependencies on external info
 - e.g. DTD, Schema



End of the archive

- Need to hand things on
 - “the bits”
 - and much else besides
- Risk losing:
 - linkage of files to experimental programs
 - ability to operate specialised programs
 - ability to link data file to system files e.g. log files





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 of this type of material not being added to the archive.





Retirement of key personnel affecting knowledge base

Lose ability to

- understand linkages between archive holdings
- locate key support texts
- maintain Ingest process
- maintain other links which have not been formally recorded





Authenticity

- Traditionally archivists have been extremely concerned about authenticity
 - Central to any “archivist” discussion about preservation
- InterPARES project:
 - “When we refer to an electronic record, we consider it essentially intact and uncorrupted if the message that it is meant to communicate in order to achieve its purpose is unaltered.”
 - Document, business process oriented
 - Does this apply to data? **What is “the” message intent of data?**
- Maintaining authenticity
 - Technical aspects **hash codes etc**
 - Social aspects **do I trust him/her?**
- Perhaps not so important for science data in the short term but will become increasingly important over time





Trusted Digital Repositories

- Invited group, hosted by Research Library Group (RLG)
- Concerned with organisational and financial issues
- Trusted Digital Repositories: Attributes and Responsibilities (TDR)
 - <http://www.rlg.org/legacy/longterm/repositories.pdf>





RLG/NARA Working Group

- International group of individuals selected by RLG/NARA
- Representatives of TDR document
- OAIS standard
- Various types of archive
- Combine
 - TDR – financial, organisational infrastructure
 - OAIS – technical issues
- Produced “***Trustworthy Repositories Audit & Certification: Criteria and Checklist***” (TRAC)



Outline of TRAC (1): Organisational Infrastructure

- A1. Governance and organizational viability
- A2. Organizational structure and staffing
- A3. Procedural accountability and policy framework
- A4. Financial sustainability
- A5. Contracts, licenses, and liabilities

Organizational infrastructure includes but is not restricted to these elements:

- Governance
- Organizational structure
- Mandate or purpose
- Scope
- Roles and responsibilities
- Policy framework
- Funding system
- Financial issues, including assets
- Contracts, licenses, and liabilities
- Transparency





Outline of TRAC (2): Digital Object Management

- B1: Ingest: acquisition of content:
 - The initial phase of ingest that addresses acquisition of digital content.
- B2: Ingest: creation of the archivable package:
 - The final phase of ingest that places the acquired digital content into the forms, often referred to as Archival Information Packages (AIPs), used by the repository for long-term preservation.
- B3: Preservation planning
 - Current, sound, and documented preservation strategies along with mechanisms to keep them up to date in the face of changing technical environments.
- B4: Archival storage & preservation/maintenance of AIPs
 - Minimal conditions for performing long-term preservation of AIPs.
- B5: Information management
 - Minimal-level metadata to allow digital objects to be located and managed within the system.
- B6: Access management
 - The repository's ability to produce and disseminate accurate, authentic versions of the digital objects.





Outline of TRAC (3): Technologies, Technical Infrastructure, & Security

- C1: General system infrastructure requirements.
- C2: Appropriate technologies, building on the system infrastructure requirements, with additional criteria specifying the use technologies and strategies appropriate to the repository's designated community(ies).
- C3: Security—from IT systems, such as servers, firewalls, or routers to fire protection systems and flood detection to systems that involve actions by people



Critique of TRAC

- Closed process
 - Single review of draft document
- Many changes based on unpublished “test audits”
- Underplays “understandability”
 - Important for data
 - Assumed not to be important for “documents”
- Simple list –
 - Do ALL boxes have to be ticked?
 - What does a “tick” mean anyway?
- Link to other standards
 - ISO 17799/27001 for security (overlap with TRAC section C)
 - ISO 9000 – say what you do and do what you say
 - but impractical to demand multiple independent audits





ISO process status

- New group set up with the primary aim of producing an ISO standard
 - Repository Audit and Certification (RAC)
- OPEN process
 - Wiki open to all
 - Mailing list open to all
 - Virtual meetings normally every week
 - See <http://wiki.digitalrepositoryauditandcertification.org>
- Into ISO via CCSDS – same route as OAIS
 - Some organisational/procedural changes in CCSDS
- Currently a Birds of a Feather (BoF) group
 - To demonstrate adequate support for the work
- Subsequently should become a Working Group
- Documents agreed by the WG will then be reviewed by CCSDS and more broadly via international ISO review process



Current status

- Reviewing and comparing
 - TRAC
 - NESTOR
 - DCC documents
- Do we need another ISO standard?
 - Could we could simply add to existing standards e.g. ISO 27001
 - The view is that ISO 27001 CANNOT be modified adequately
 - It's view of Information is too limited
- Started drafting a straw man document
 - Taking TRAC and add concepts from other docs





Key Issues

- How to get from a checklist to an international accreditation/ certification system?
- Evidence – short term
- Evidence – long term
 - The real crunch!
- Quantification
 - The marking system
- Levels of audit?
 - External review
 - Internal maturity





The Market

- Transparency
- Trustable?
 - certified by whom?
 - to what level?
 - what evidence?
 - for what Designated Community
 - relevant/sensible?
- What cost?



Links

- RAC group Wiki:
 - <http://wiki.digitalrepositoryauditandcertification.org>
- TRAC document
 - <http://www.crl.edu/PDF/trac.pdf>
- Digital Curation Centre
 - <http://www.dcc.ac.uk>
- 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>

