International Audit and certification of Digital Repositories

David Giaretta and Simon Lambert
STFC, Rutherford Appleton Laboratory, Didcot, Oxon OX11 0QX, UK
Email: david.giaretta@stfc.ac.uk

ABSTRACT

There are a large number of repositories which claim to be able to preserve digitally encoded information. Funders invest a great deal of resources in such repositories; depositors entrust their valuable data to them. How can either have any confidence that the repositories can be trusted?
While there is a natural reluctance on the part of many repositories to be independently tested, nevertheless there has been a longstanding demand for a way to identify repositories which could be trusted. How can these demands be satisfied and how can any confidence be placed in any such judgement, and indeed is it possible to obtain any consistency in such a judgment? These are not clearly matters which are amenable to simply a technical solution nor are they matters which are easily checked.
This paper will provide some background for the lengthy efforts to produce an ISO standard which can form the basis of an international accreditation and certification process. Beside the fundamental principles on which metrics are based, this paper will describe the procedural, social and legal framework being created to support this effort and provide an up to date status report.

INTRODUCTION

The Preserving Digital Information report of the Task Force on Archiving of Digital Information [1] 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.

The issue of certification, and how to evaluate trust into the future, as opposed to a relatively temporary trust which may be more simply tested, has been a recurring request, repeated in many subsequent studies and workshops.
The OAIS Reference Model Open Archival Information System (OAIS) [7], is now adopted as the "de facto" standard for building digital archives [6]. Section 1.5 of OAIS (Road map for development of related standards) included an item for accreditation of archives, reflecting the long-standing demand for a standard against which Repositories of digital information may be audited and on which an international accreditation and certification process may be based. It was agreed that RLG and NARA take a lead on this follow-on standard. This they did, forming a closed panel which produced Trustworthy Repositories Audit & Certification: Criteria and Checklist [10].
TRAC was based on two documents, namely the OAIS Reference Model [7] and the Report on Trusted Digital Repositories: Attributes and Responsibilities [8]. The former lays out fundamental requirements for preservation, while the latter focussed on the administrative, financial and organisational requirements for the body undertaking the preservation activities.
Other, separate, work includes the nestor Catalogue of Criteria for Trusted Digital Long-term Repositories [5], which is also based on OAIS. The next section explains the advantages of OAIS for approaching certification of digital repositories.

TESTABILITY AND KEY OAIS CONCEPTS

As a precursor to discussing its preservation, one may begin by asking what the definitions of information or data might be - how restrictive do we need to be? OAIS provides a very general definition of Information, namely: *Any type of knowledge that can be exchanged. In an exchange, it is represented by data.*
Information clearly includes data as well as documents, and covers behaviour, performance and explicit, implicit and tacit information.
Data is defined as: *A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing*.

Preservation

We need first some methodology by which to test the basic claim that someone is preserving some digitally encoded information; without such a test this is a meaningless claim. OAIS introduces the, quite reasonable, test that the digital

object must somehow be useable and understandable in the future. However by itself this is too broad - are we to be forced to ensure that the digitally encoded designs of a battleship are to be understood by everyone, for example a 6 year old child? In order to make this a practical test the obvious next refinement is to describe the type of person - and more particularly their background knowledge - by whom the information should be understandable. Thus OAIS introduces the concept of **Designated Community**, defined as 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. Note that a Designated Community is defined by the repository and this definition may change/evolve over time.

Bringing these ideas together we can then say, following OAIS, that preserving digitally encoded information means that we must ensure that the information to be preserved is **Independently Understandable** to (and usable by) the Designated Community.

We are clearly concerned about long term preservation, but how long is that? OAIS defines **Long Term** as long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing Designated Community. Long Term may extend indefinitely

OAIS contains a number of models. The most important of these is the Information Model, shown in Figure 1.

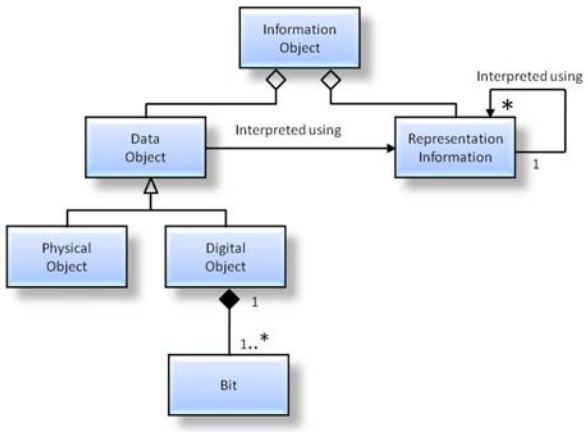

Figure 1 OAIS Information Model

Representation Information

The UML diagram (Figure 1) means that

- an Information Object is made up of a Data Object and Representation Information
- a Data Object can be either a Physical Object or a Digital Object. An example of the former is a piece of paper or a rock sample
- a Digital Object is made up of one or more Bits
- a Data Object is interpreted using Representation Information
- Representation Information is itself interpreted using further Representation Information

The figure shows that Representation Information may contain references to other Representation Information. When this is coupled with the fact that Representation Information is an Information Object that may have its own Digital Object and other Representation Information associated with understanding each Data Object, as shown in a compact form by the “interpreted using” association loop, the resulting set of objects can be referred to as a Representation Network.

A Representation Network should cover semantic and structural information as well as recognising that there may be Other Representation Information such as software.

The recursion in Representation Information leads to the question of how and where this recursion ends. Given the definitions one can see that the natural end of the recursion lies with what the Designated Community knows i.e. the Knowledge Base, defined as a set of information, incorporated by a person or system, that allows that person or system to understand received information, of the Designated Community. Once again, experience shows that any such Knowledge Bases changes over time, the changes ranging from the introduction of new theories to drift in vocabularies.

Definition of the Designated Community

An important clarification is needed here, namely that the definition of the Designated Community is left to the preserver. The same digital object held in different repositories could be being preserved for different Designated Communities, each of which could consist of many disjoint communities.
The quid pro quo is that those funding or entrusting digital objects to the repository can judge whether the definition of the Designated Community is appropriate for their needs.

OAIS CONFORMANCE

OAIS defines a number of responsibilities by which to judge conformance, which may be summarised as an OAIS must (these are the likely revised versions of these responsibilities)

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the archive, ensuring that it is never deleted unless allowed as part of an approved strategy - there should be no ad-hoc deletions,
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Object., with evidence supporting its Authenticity.

OAIS introduces a number of important concepts and conformance criteria; however this is not enough on which to base a certification scheme. The next section describes some of the factors which must also be taken into consideration.

WHAT CAN CHANGE?

We can consider some of the things can change over time and hence against which an archive must safeguard the digitally encoded information.

Hardware and Software Changes

Use of many digital objects relies on specific software and hardware, for example applications which run on specific versions of Microsoft Windows which in turn runs on Intel processors. Experience shows that while it may be possible to keep hardware and software available for some time after it has become obsolete, it is not a practical proposition into the indefinite future, however there are several projects and proposals which aim to emulate hardware systems and hence run software systems.

Environment Changes

These include changes to licences or copyright and changes to organisations, affecting the usability of digital objects. External information, ranging from the DNS to DTDs and Schema, vital to the use and understandability, may also become unavailable.

Termination of the Archive

Without permanent funding, any archive will, at some time, end. It is therefore possible for the bits to be lost, and much else besides, including the knowledge of the curators of the information encoded in those bits. Experience shows that much essential knowledge, such as the linkage between holdings, operation of specialised hardware and software and links of data files to events recorded in system logs, is held by such curators but not encoded for exchange or preservation. Bearing these things in mind it is clear that any repository must be prepared to hand over its holding – together with all these tacit pieces of information – to its successor(s).

Changes in what people know

As described earlier the Knowledge Base of the Designated Community determines the amount of Representation Information which must be available. This Knowledge Base changes over time.

AUTHENTICITY

Trustability of holdings involves being sure of the authenticity of such holdings. Much has been written about authenticity and its role in preservation, for example in the InterPARES project (http://www.interpares.org/). While it seems unreasonable to require all archives themselves to investigate the origins of their holdings, it is reasonable for archives to be able to maintain authenticity. To maintain authenticity, evidence must cover both technical aspects, using techniques such as digests and hashes which can be used to prove that the bit sequences have not been changed unexpectedly, and social aspects, namely who has been entrusted with what, for example computer system administrators.

If the digital object is transformed then the bit sequences will have been changed. In this case there must be some evidence that the information encoded in the digital object is unchanged. In order to do this a number of tests may be performed by appropriate, and one hopes, trustworthy, people. For example the data values of a scientific dataset before and after the transformation may be compared and verified as the same. For digital objects which are normally rendered, for example PDF files or JPEG images, then there might be other tests, often called **Significant Properties**, which can be evaluated and verified as unchanged after the transformation. A full discussion of this topic is outside the scope of this paper.

TRAC AND RELATED DOCUMENTS

A group was gathered by NARA and RLG (the latter subsequently incorporated into OCLC) to form the Task Force on Trusted Digital Repositories. This group produced the Trustworthy Repositories Audit and Certification : Criteria and Checklist [10]. The work combined concepts from OAIS and the Trusted Digital Repositories: Attributes and Responsibilities6 [8]. The latter allowed the group to supplement OAIS with considerations of financial stability and training of personnel.

The document has a number of metrics grouped into

- Organisational Infrastructure
- Digital Object Management and Technologies
- Technical Infrastructure
- Security.

Accompanying each of the metrics is extensive additional explanatory text and examples of the types of evidence which might be used as proof of fulfilling the metrics. The document has being used as the basis for internal and test audits in a number of repositories, however it is not part of a formal audit and certification process.

Other work in this area includes:

- the German preservation consortium, nestor, has produced a Catalogue of Criteria for Trusted Digital Repositories [5]
- in early 2007 representatives from the Digital Curation Center (DCC, http://www.dcc.ac.uk), DigitalPreservationEurope (DPE, http://www.digitalpreservationeurope.eu/), NESTOR (Germany) and the Center for Research Libraries (North America) met and produced a list of 10 core criteria for digital preservation repositories, to guide further international efforts on auditing and certifying repositories [2]. A comparison of this list with the OAIS responsibilities was produced [4].
- Ross et al [9] produced comments on the TRAC document
- a cross-walk between the TRAC, nestor and Ross documents was produced [3]
- the DCC and DPE projects produced the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) toolkit. This toolkit is intended to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognise their strengths.

All this work has been helpful in providing information and experience in assessing digital repositories, and some provide a local or project-backed certificate of quality. However none provide an ISO based accreditation and certification system of the kind which are available in other areas, such as the one concerning Information Security, based on ISO 27001 series. Without this we cannot expect to have a mark of quality and trustability for digital repositories which is recognised world-wide. Efforts to produce such a system are described next.

DEVELOPMENT OF AN ISO ACCREDITATION AND CERTIFICATION PROCESS

The development of OAIS was hosted by the Consultative Committee for Space Data Systems (CCSDS, http://www.ccsds.org) and approved by ISO as ISO 14721. OAIS contained a roadmap which listed a number of possible follow-on standards, some of which e.g. the Producer-archive interface -- Methodology abstract standard (ISO 20652:2008), have already become ISO standards, after development within CCSDS.

The need for a standard for certification of archives was included in that list and the RLG/NARA work which produced TRAC was the first step in that process. The next step was to bring the output of the RLG/NARA working group back into CCSDS. This has been done and the Digital Repository Audit and Certification (RAC) Working Group [11] has been created, the CCSDS details are available from http://cwe.ccsds.org/moims/default.aspx#_MOIMS-RAC, while the working documents are available from http://wiki.digitalrepositoryauditandcertification.org. Both may be read by anybody but, in order to avoid hackers, only authorised users may add to them. The openness of the development process is particularly important and the latter site contains the notes from the weekly virtual meetings as well as the live working version of the draft standards.

Besides developing the metrics, which started from the TRAC document, the working group also has been working on the strategy for creating the accreditation and certification process. Review of existing systems which have accreditation and certification standard processes it became clear that there was a need for two documents

1. **Audit and Certification of TrustworthyDigital Repositories**
2. **Requirements for Bodies Providing Audit and Certification of Trusted Digital Repositories.**

The first document lists the metrics against which a digital repository may be judged. It is anticipated that this list will be used for internal metrics or peer-review of repositories, as well as for the formal ISO audit process. In addition tools such as DRAMBORA could use these metrics as guidance for its risk assessments.

It must be recognised that the audit process cannot be specified in very fine, rigid, detail. An audit process must depend upon the experience and expertise of the auditors. For this reason the second document sets out the system under which the audit process is carried out; in particular the expertise of the auditors and the qualification which they should have is specified. In this way the document specifies how auditors are accredited and thereby helps to guarantee the consistency of the audit and certification process. For this reason the RAC Working Group refers to accreditation and certification processes.

At the time of writing both documents are in an advanced state of preparation; the first of these documents has been submitted for ISO review and the second should be submitted soon. It is hoped that the review process and in the Spring of 2009. While the reviews are underway further preparations for the accreditation and certification processes will be undertaken. It should be noted that the OAIS reference Model has also been undergoing revision and the new version has been submitted for ISO review. Because of the close links between the metrics and OAIS concepts and terminology it is important that the two remain consistent, and cross-membership of the working groups will ensure this.

In addition to the “central” accreditation body there will be an eventual need for a network of local accreditation and certification bodies.

CONCLUSIONS

It has been recognised for a long time that there is a need for a way to judge the extent to which an archive can be trusted to preserve digitally encoded information. On the one hand funders of such archives need some formal certification process to provide assurance that their funding is well spent and that their important digital holdings will continue to be usable and understandable into the future. On the other hand it is probably also true that many who manage such archives would want some less formal process.

Considerable work has been carried out on the second of these aims, namely peer or informal certification. The RAC Working Group seems, at the time of writing, to be close to take important steps towards the first aim (formal ISO certification). Difficult organisational issues still need to be addressed but there is a clear roadmap for doing this. Even if all this is put in place the take-up of the process and its impact on, for example, determining the funding of digital repositories is far from guaranteed. However in order to make progress the RAC Working Group believes that the effort must be made.

REFERENCES

[1] Garrett, J. & Waters, D, (Eds). (1996). Preserving Digital Information, Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group. Retrieved from http://www.ifla.org/documents/libraries/net/tfadi-fr.pdf

[2] CRL,(2007), Retrieved from http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92

[3] Dale, R., (2007), Mapping of Audit & Certification Criteria for CRL Meeting (15-16 January 2007). Retrieved from http://wiki.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/TRAC-Nestor-DCC-criteria_mapping.doc

[4] Giaretta, D., (2008), Comparison of OAIS and the Chicago Meeting 10 points. Retrieved from http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/ComparisonOaisAndChicago10Points

[5] nestor Working Group Trusted Repositories – Certification, (2006), Catalogue of Criteria for Trusted Digital Repositories. English version retrieved from http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf

[6] National Science Foundation Cyberinfrastructure Council (NSF, 2007), Cyberinfrastructure Vision for 21st Century Discovery. Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf
[7] Open Archival Information System (OAIS) – Reference Model, ISO 14721:2003, (2003). Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf
[8] RLG-OCLC, (2002), Report on Trusted Digital Repositories: Attributes and Responsibilities. Retrieved from http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf
[9] Ross, S., Bütikofer, N., and McHugh, A. (2006), DCC Comments on RLG/NARA Audit and Certification Checklist. Retrieved from http://wiki.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/Ross_McHugh_Buetikofer_comments_RLGNARA_AUDIT_ver2.pdf
[10] TRAC, (2007), Trustworthy Repositories Audit & Certification: Criteria and Checklist. Retrieved from http://www.crl.edu/PDF/trac.pdf
[11] Repository Audit and Certification Working http://wiki.digitalrepositoryauditandcertification.org

ACKNOWLEDGEMENTS

The entire RAC Working Group who deserve credit for the vision and effort which they are putting into this work.