

A study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation

eBank-UK Phase 3: WP4
September 2006 - June 2007

Final Version (Revised): 7th September 2007

Manjula Patel
UKOLN, DCC
University of Bath, UK

Simon Coles
National Crystallography Centre
University of Southampton, UK



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 2.0 UK:
England & Wales License](https://creativecommons.org/licenses/by-nc-sa/2.0/uk/).

Revision History

Date	Contributor	Revision
21 st May 2007	Manjula Patel	Outline structure of report; audit and certification; representation information; preservation metadata; eprints.org platform
22 nd May 2007	Manjula Patel	Scoping of additional subsections under each section heading
23 rd May 2007	Manjula Patel	Started to add text into sections on audit and certification
28 th May 2007	Manjula Patel	Revision of structure and text for discussion at F2F meeting with Simon Coles on 29 th May 2007.
1 st June 2007	Manjula Patel	Incorporation of changes from F2F meeting
4-6 th June 2007	Manjula Patel	Completion of audit and certification sections
7 th June 2007	Manjula Patel	Recommendations for audit and certification section
8 th June 2007	Manjula Patel	Started sections on OAIS Model and representation information
12-13 th June 2007	Manjula Patel	Completion of section 3
19-22 nd June 2007	Manjula Patel	Text for section on preservation metadata (section 4) and tidying up of all text for discussion at eBank Phase 3 project meeting on 25 th June.
26-27 th June 2007	Manjula Patel	Tidying up of text and first draft of ePrints.org section
28 th June 2007	Manjula Patel	Circulated to Simon Coles and Liz Lyon for comment
29 th June 2007	Manjula Patel	Addition of recommended PREMIS Semantic Units
7 th July 2007	Simon Coles	Comments and modifications to various sections
16 th July 2007	Manjula Patel	Continued refinement of text and recommendations based on feedback from Simon Coles. Circulated report for comment to Simon Coles, Liz Lyon, Les Carr, Jeremy Frey, Monica Duke and ukoln-dcc mailing list.
20 th July 2007	Manjula Patel	Tidying up of text and typographical errors pointed out by Alex Ball
27 th July 2007	Manjula Patel	Deadline for comments and feedback reached. Refined parts of text and circulated to Liz Lyon and Simon Coles
30 th July 2007	Manjula Patel	Final Version; added CC license and revision history
6-7 th September 2007	Manjula Patel	Added summary and tidied up recommendations and conclusions as suggested by Steve Hitchcock. Some reformatting of text and page breaks.

Contents

1. Introduction	5
2. Audit, Certification and Trustworthiness	6
2.1 Current Audit and Certification Instruments	7
2.1.1 Trustworthy Repositories Audit and Certification Checklist (TRAC).....	8
2.1.2 Network of Expertise in Long-Term Storage of Digital Resources (NESTOR)	9
2.1.3 Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)	9
2.1.4 International Repository Audit and Certification Birds of a Feather Group	11
2.2 Auditing the eCrystals Data Repository	11
2.4 Recommendations.....	12
3. The Open Archival Information System (OAIS)	13
3.1 Important Concepts.....	14
3.2 Functional Model.....	14
3.4 Information Model.....	15
3.5 Representation Information and Networks	16
3.7 The DCC Registry/Repository of Representation Information (RRoRI).....	17
3.7.1 Types of Representation Information.....	18
3.8 Representation Information in eCrystals	19
3.8.1 Populating the RRoRI with eCrystals Data RI.....	20
3.9 Recommendations.....	22
4. eBank-UK Application Profile and Preservation Metadata	22
4.1 OAIS Preservation Description Information.....	23
4.2 PREMIS Data Dictionary	24
4.3 Preservation Metadata in eCrystals.....	26
4.4 Recommendations.....	27
5. ePrints.org Conformance to OAIS.....	28
6. Conclusions	29
Acknowledgements.....	29
References	29
Appendix I –TRAC Audit Checklist.....	33
Appendix II–DRAMBORA Checklist.....	34

Summary

The JISC funded eBank-UK project (two phases since Sept. 2003) [1], has investigated the linking of primary data to other research outputs, such as published papers, within the scholarly knowledge cycle [2]. Building on the concept of open access [3], the project has focussed on the laboratory based experimental technique of chemical crystallography at the UK National Crystallography Centre. It has constructed an institutional data repository (eCrystals) that makes available the raw, derived and results data from a crystallographic experiment [4]. Following the creation of a completed crystal structure, data is uploaded into a data repository and additional metadata (chemical as well as bibliographic) is associated with the dataset.

The objectives of eBank-UK: Phase 3 (Sept. 2006–June 2007) [5] aim to progress the establishment of a global federation of data repositories for crystallography: the “eCrystals Federation”, through a comprehensive feasibility scoping study. An important part of the study (this report) explores data curation and preservation issues as well as sustainability within a federation. Long-term sustainability of digital data requires, in the first instance, a policy commitment to undertake curation and preservation duties in maintaining the data so that it is usable (and reusable) for its useful lifetime. However, such a commitment is likely to be influenced by a whole host of factors including social, political, organisational, financial and technical. One way of assessing these factors in the context of the eCrystals data repository and federation is to consider the questions posed in the rapidly developing area of repository audit and certification.

Within the preservation community, the *Reference Model for an Open Archival Information System* (OAIS) (ISO 14721:2003 [14]), has established itself as an important standard, influencing: the development of preservation metadata; architectures and systems design of repositories; and conformance criteria for archival repositories. Although the OAIS standard covers a wide range of issues relating to the operating environment of an archive or repository, its concept of using *Representation Information* (RI) as a means of preserving access to the information content of digital objects is currently receiving significant attention [32, 34, 35]. Consequently, we devote a section to examining the RI of the content held in the eCrystals repository, in particular the variety of file formats in use.

Fundamental to preserving and curating digital information, is the recording of adequate and appropriate metadata. Whilst the exact metadata to be recorded is dependent on the specific preservation strategy in force, there is some consensus on a certain core set of preservation metadata (PREMIS Data Dictionary [53]). We therefore examine the implications of this on the eBank-UK Metadata Application Profile [46].

The capabilities and constraints of the software platform underlying a repository are critical to the functions and services that can be provided at the application level. With this in mind, we take a brief look at the ePrints.org software upon which the eCrystals data repository is constructed.

Given the exploratory nature of this report, we have tried to identify issues that are likely to impact on the long-term preservation, curation, maintenance and sustainability of crystallography data and in particular the eCrystals data repository. In order to progress this work and take it forward in the context of a federation we follow each major topic area with a set of recommendations, some of which have over-lapping scope.

1. Introduction

The JISC funded eBank-UK project (two phases since Sept. 2003) [1], has investigated the linking of primary data to other research outputs, such as published papers, within the scholarly knowledge cycle [2]. Building on the concept of open access [3], the project has focussed on the laboratory based experimental technique of chemical crystallography and constructed an institutional data repository (eCrystals) that makes available the raw, derived and results data from a crystallographic experiment [4]. Following the creation of a completed crystal structure, data is uploaded into a data repository and additional metadata (chemical as well as bibliographic) is associated with the dataset. This approach allows rapid release of crystal structure data into the public domain, but can also provide mechanisms for value added services that allow discovery of the data for further studies and reuse, whilst ownership of the data is retained by the creator.

The objectives of eBank-UK: Phase 3 (Sept. 2006–June 2007) [5] aim to progress the establishment of a global federation of data repositories for crystallography: the “eCrystals Federation”, through a comprehensive feasibility scoping study which aims to: assess organisational issues and promote advocacy; examine interoperability associated with research workflow and data deposit; harmonise the metadata application profiles from repositories operating on different platforms; investigate aggregation issues and scope federation relationships with an international subject archive (based at the International Union of Crystallography (IUCr) [6]) and the Crystal Structure Database (CSD) at the Cambridge Crystallographic Data Centre (CCDC) [7]. The eBank-UK aggregator service [1] will further the issues related to linking datasets with primary sources of publication.

In addition, Phase 3 will also explore data curation and preservation issues as well as sustainability within a federation. Long-term sustainability of digital data requires, in the first instance, a policy commitment to undertake curation and preservation duties in maintaining the data so that it is usable (and reusable) for its useful lifetime. However, such a commitment is likely to be influenced by a whole host of factors including social, political, organisational, financial and technical. One way of assessing these factors in the context of the eCrystals data repository and federation is to consider the questions posed in the rapidly developing area of repository audit and certification.

According to the DCC [8]

“Digital curation is maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific materials.”

The e-Science Curation Report (2003) by Lord and Macdonald [9] proposed the following distinctions:

Curation: The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.

Archiving: A curation activity, which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.

Preservation: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

Furthermore, the term “digital curation” is increasingly being used for the actions needed to maintain and utilise digital data and research results over their entire life cycle for current and future generations of users (see JISC circular 6/03 (Revised) [10]) including using information in new ways and also publishing results based on the information.

A recent meeting of leading experts actively working on the audit and certification of preservation repositories identified the following ten desirable characteristics of long-term digital repositories [11]:

1. The repository commits to continuing maintenance of digital objects for identified community/communities.
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfil its commitment.
3. Acquires and maintains requisite contractual and legal rights and fulfils responsibilities.
4. Has an effective and efficient policy framework.
5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
6. Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
7. Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.
8. Fulfils requisite dissemination requirements.
9. Has a strategic program for preservation planning and action.
10. Has technical infrastructure adequate to continuing maintenance and security of its digital objects.

In addition to developments in audit and certification issues, we have identified other aspects relevant to the curation of data in the eCrystals data repository including: the concept of Representation Information (section 3); preservation metadata in relation to the eBank Metadata Application Profile (section 4) and conformance of the software underlying the eCrystals data repository (ePrint.org [12]) to the *Open Archival Information System (OAIS) Reference Model* [13] which is the predominant reference model used in the world of digital preservation (section 5). We conclude this scoping study with recommendations under each of these sections for further developing the eCrystals repository and associated federation from the point of view of curation and preservation issues.

2. Audit, Certification and Trustworthiness

Audit and certification processes have always been recognised as an important aspect of digital preservation. Both the *Trusted Digital Repositories* report [15] and the OAIS standard (ISO 14721:2003) [14] recommended the setting up of processes for audit and certification of repositories. Their application in the context of digital preservation and curation has seen rapid evolution over the past few years.

Digital preservation is a heavy risk activity due to dynamic and unpredictable developments in technology over both the short and long term. Digital information is subject to change, modification and obsolescence—any of which can happen very easily if the data is not managed adequately. The vulnerability of digital information as well as its prolific creation [16] demand that those entrusted with its stewardship are demonstrably trustworthy to the community that they serve.

Audit and certification is one method of engendering trust in those charged with looking after digital data (most often in the form of digital repositories). To owners of digital content looking to deposit their data for long-term survival, a repository's trustworthiness will be of paramount importance. The Commission on Preservation and Access (CPA) and Research Libraries Group (RLG) Task Force on Archiving of Digital Information asserted in 1996 [17 (page 40)]:

“...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.”

However, demonstrating trust is not an easy task. In 2000, an RLG/OCLC Working Group explored networks of trust relationships in the report *Trusted Digital Repositories: Attributes and Responsibilities* [15]. They found that trust relationships are multi-faceted and dependent on many different aspects of a repository's processes and workflows. Furthermore, different stakeholders are interested in different aspects of “trustworthiness”: for example, funding bodies are interested in statistics relating to the ingest of data objects and the number and type of user visits or requests; users are concerned about the context and authenticity of data as well as added-value services whilst depositors care about intellectual property rights, preservation and the visibility of their deposited content.

The trustworthiness of a content provider depends on several things, including the expertise of the staff, the workflows and the quality control measures that are in place. The trustworthiness of the digital information itself is largely dependent on information about the data itself: what has happened to it; its origin or provenance and by whom it is being managed.

Closely related to the issue of trustworthiness is the question of what should be audited to ensure that trustworthiness can be achieved: should the audit include the whole organisation? or merely the processes through which the digital objects are managed? or the system within which the information is contained? or should we examine aspects of all of these? Establishing the scope of the audit is crucial for realising the value of the audit report and what can be achieved with it.

Audits fit into the third stage of the Deming Cycle (Plan, Do, Check, Act) [18] and can be achieved through self-assessment, or by internal or external audit. The goal is to check whether the identified objectives of the planning stage are met. An audit checks whether what has happened or is happening complies with what was originally planned or legally laid out. An audit does not directly improve the situation; it only describes and assesses it. However, the assessment should encourage action towards improvement and in some cases may provide explicit recommendations for improving the analysed situation.

Certification is basically a designation earned by a repository to assert that it has a specific set of knowledge, skills and capabilities to correctly do the job that it purports to be doing, in the view of the certifying body. Certification may be used as a means of communicating the trustworthiness of a preservation service to possible clients and in comparing repository services.

2.1 Current Audit and Certification Instruments

There is a growing movement underway to develop metrics that measure the quality or trustworthiness of a digital repository. Proponents of certification believe this process will create more standardized, reliable and credible archives that better meet the long-term needs of digital information users.

Standards to ensure the quality, authenticity, reliability and integrity of digital information have been in existence for some time. These include: the ISO 15489 records management standard which identifies the requirements of authenticity, reliability, integrity and usability for both records and records systems as well as the processes that manage them; the information security standard (ISO 17799) which provides a framework for implementing security requirements and quality management (ISO 9001).

However, digital curation and preservation transcends all of these aspects; the long-term survival of digital information in a repository is dependent on the repository organisation's financial, physical, political and cultural viability as well as the repository system's technical security and the authenticity and integrity of the data that it holds. For this reason, several groups of experts have been actively moving forward the audit and certification agenda. Increasingly these groups are working collaboratively and on an international basis.

2.1.1 Trustworthy Repositories Audit and Certification Checklist (TRAC)

This effort to develop criteria for trustworthy digital repositories began in 2002 with the publication of the RLG-OCLC report entitled *Trusted Digital Repositories: Attributes and Responsibilities* [15]. The report defined: the characteristics of a trusted digital repository; listed relevant attributes of such a repository; called for compliance with the OAIS as well as administrative responsibility, organisational viability, financial sustainability, technological and procedural suitability, system security and procedural accountability. It also recommended that a process be developed for the certification of digital repositories.

Based on this foundational work, in 2003, the RLG-NARA Digital Repository Certification Task Force was established to develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections. The international task force produced a set of certification criteria applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services. This checklist was made available in the form of a draft for public comment in 2005 [19].

Also in 2005, the Andrew W. Mellon Foundation awarded funding to the Center for Research Libraries (CRL) to: further establish the documentation requirements; delineate a process for certification; and establish appropriate methodologies for determining the soundness and sustainability of digital repositories. Leveraging the audit checklist developed by RLG and NARA [19], a CRL project (Auditing and Certification of Digital Archives) conducted several pilot audits (including the E-Depot at the Koninklijke Bibliotheek in the Netherlands, the Inter-University Consortium for Political and Social Research, and Portico), informing further checklist development. Findings and methodologies were shared with those of related working groups in Europe who applied the draft checklist in their own domains: the Digital Curation Center (U.K.) [8], Digital Preservation Europe (Continental Europe) [20] and NESTOR (Germany) [21], culminating in the latest checklist, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, (TRAC) published in March 2007 [22]. TRAC is now under the stewardship of the CRL, which will oversee its further development.

TRAC takes the OAIS as its intellectual foundation and splits the audit criteria into three categories: organisational infrastructure; digital object management and finally technologies, technical infrastructure and security. The major difference between the RLG-NARA checklist and TRAC is the requirement for documentary evidence relating to various issues, in particular policy and sustainability. The 93-page report also provides considerable advice on the application of the checklist.

2.1.2 Network of Expertise in Long-Term Storage of Digital Resources (NESTOR)

In parallel to the work being undertaken by RLG, NARA and CRL (see section 2.1.1), Germany's NESTOR (Network of Expertise in Long-Term Storage of Digital Resources) project has been aiming to set up "criteria for trusted digital repositories [and] recommendations for certification procedures of digital repositories." The NESTOR catalogue comprises 14 criteria, grouped into sections entitled: organisational framework, object management and, infrastructure and security. The basic principles for determining trustworthiness using the NESTOR catalogue are: measurability in terms of quantification; documentation to provide evidence and transparency to promote openness.

In addition to the NESTOR catalogue, the DINI (Deutsche Initiative für Netzwerkinformation) Workgroup on Electronic Publishing released its process for the certification of institutional repositories in 2003, with the English language version appearing in 2006 [23,24]. This document is part of the policy framework of the German initiative for network information. It establishes minimum requirements for institutional repositories in the higher education sector. The goal is to facilitate the construction of a converging infrastructure that will ultimately allow the open exchange of scientific publications across Germany.

Criteria of the DINI 2004 certificate embrace policy issues, legal aspects and communication with information producers, as well as interfaces for dissemination. With regard to digital preservation they demand documentation standards and a minimum retention term of five years. Optional preservation criteria include OAIS compliance and references to further recommendations. Certification of a repository entails audit by a workgroup, the procedure takes approximately two months. If certification is successful the audited institutional repository is allowed to display the logo of the DINI certificate with the date of issue on their website.

While this certificate does not focus on long-term digital preservation, it constitutes a basic infrastructure upon which comprehensive audit and certification in digital preservation can be built. The DINI certificate is currently the only audit framework that addresses digital preservation issues resulting in certification.

2.1.3 Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)

The DRAMBORA toolkit is the latest audit instrument to be made publicly available in March 2007 [25]. It stems from discussions by the DCC and DPE with: the TRAC working group; the CRL Certification of Digital Archives Project; NESTOR and the international repository audit and certification birds of a feather group (see section 2.1.4).

The toolkit results from the experience gained through undertaking pilot audits by the DCC based on the RLG-NARA checklist and the NESTOR catalogue. It places emphasis on documented evidence and the assessment and management of risks as critical factors in determining the trustworthiness of a repository [26, 27]. Digital curation is seen to be about taking organisational, procedural, technological and any other uncertainties and transforming them into manageable risks. It attempts to pose questions such as:

- is a repository capable of identifying and prioritising the risks that may impede its activities?
- is a repository managing the risks to mitigate the likelihood of their occurrence?
- is a repository establishing effective contingencies to alleviate the effects of the risks that may occur?

Furthermore, the toolkit was designed specifically with the aim of self-audit in mind. It recognises that an authentic and understandable digital object is at the centre of the audit process—can the repository be trusted to deliver the authentic and understandable digital object to the end user and over what period of time?

The toolkit draws on existing work in the area of enterprise risk management, which is based on identifying the context within which risks need to be managed; the risks themselves; assessing and evaluating risks and defining measures to address and manage those risks. Consequently, DRAMBORA requires auditors to undertake the following six stages during the audit process:

1. Identification of objectives (business context)
2. Identification of policy and regulatory framework
3. Identification of activities and assets
4. Identifying risks related to activities and assets
5. Assessing risks
6. Managing risks

As part of the audit, the following 10 questions are posed [28 (slides 30-31)]:

1. What is the mandate of your repository?
2. What are the goals and objectives of your repository?
3. What policies does your repository have in place to support and regulate how these goals and objectives are to be achieved?
4. What legal, contractual and other regulatory requirements/confines does your repository operate in?
5. What standards and codes of practice does your repository follow?
6. Any other things that influence how your repository does whatever it is supposed to be doing?
7. What activities does your repository undertake to achieve its goals and objectives within the context and confines set by the regulatory environment, and what assets do you use and produce in the course of these activities, including staff, skills, knowledge, technology?
8. What are the risks associated with all of the above?
9. How would you assess these risks?
10. How do you manage these risks?

The self-audit produces a composite risk score for each of eight functional classes, grouped into two types:

- Organisational: acquisition and ingest, storage and preservation, metadata management, access and dissemination
- Support: organisation and management, staffing, financial management, technological solutions and security

These numeric risk scores allow the identification of areas that are most vulnerable to threats. However, auditors should be aware that there may be inter-relationships that exist between specific risks.

Following a risk audit, it is expected that the following will have been achieved:

- A comprehensive and documented self-awareness of the repository's mission, aims and objectives as well as of the activities and assets intrinsic to these
- Construction of a detailed catalogue of pertinent risks, categorised according to type and inter-risk relationships, including the probability and potential impact of each risk
- An internal understanding of the successes and shortcomings of the organisation
- Preparation of the organisation for subsequent external audit

The DRAMBORA toolkit is newly released and as such, the DCC and DPE are planning test audits to gain experience and feedback in its use with the intention of releasing version 2.0 in

Sept. 2007 in the form of an on-line interactive toolkit. The plan is to release a further version in spring 2008.

2.1.4 International Repository Audit and Certification Birds of a Feather Group

This working group was recently set up, with its first meeting taking place in January 2007. It has established a Wiki, which contains information and documents generated by the working group [29]. The aim of the group is to

“...produce an ISO standard on which a full audit and certification of digital repositories can be based. The aim will be to take this work into ISO in the same way as the OAIS Reference Model (ISO 14721), namely via ISO TC20/SC13, of which the working arm is CCSDS.”

The membership of the group is self-nominating. Minutes of the discussions and intermediate drafts and working papers will be available for public scrutiny and comment.

Section 1.5 of the OAIS Reference Model (ROAD MAP FOR DEVELOPMENT OF RELATED STANDARDS) included an item *standard(s) 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. The RLG-NARA work forms input to the BOF WG. Other inputs are expected, including the NESTOR Catalogue of Criteria for the ongoing changes in the business environment.

2.2 Auditing the eCrystals Data Repository

The eCrystals data repository started life as a prototype research data repository with the over-arching aim of sharing and disseminating data within the crystallography domain. In the same manner as other University research repositories it is characterised by short-term staffing contracts and research funding cycles. Nevertheless, it is currently in the process of maturing into a valuable community resource. At present it has no formal long-term commitment to preservation, but it is clearly becoming an invaluable asset to the crystallography community.

Whilst the audit process can be used for varying reasons, for example to validate the processes and procedures of a particular repository or to prepare a repository for a subsequent formal certification process, it also has merits in providing input into the development stages of a repository. In the case of the eCrystals repository, there is an element of each of these objectives. The National Crystallography Service (NCS) and the eBank-UK project are seeking to validate the processes already in place, but also looking to audit the eCrystals repository for risk-assessment purposes as well as to aid in the further evolution of the repository to better serve its “designated community”[13]. It should be recognised that given the rapidly changing nature of the technologies in use, it is important to regularly monitor the environment in which the repository operates. An effective way of achieving this aim is to undertake audits on a regular basis so that a profile of the repository and its operating environment can be built up over time.

As part of WP4 in Phase 3 of the eBank-UK Project, we began by investigating the RLG/NARA checklist, since this was the most complete audit instrument at the start of Phase 3 (the NESTOR catalogue was published in the German language and has only recently become available in the English language).

The RLG/NARA checklist was found to be very large, “heavy” and comprehensive; it appeared to be more suited to large national archive services and their progression to formal

certification than to a research repository. It is clear that the eventual aim of this work has always been to develop an audit process that results in certification. It has been designed to help institutions objectively evaluate responsibilities against capabilities and identify potential risks to digital content held in repositories and other archives.

Once the NESTOR catalogue became available in an English language version, we also examined the NESTOR and DINI work. This appeared to have similar aims to the RLG/NARA work, also being based on the OAIS Reference Model, but it is largely geared to a German national context.

More recent developments reflect an emphasis on proving trustworthiness based on documentary evidence (TRAC) and risk-assessment (DRAMBORA). A major difference between the RLG/NARA checklist and TRAC is that documented evidence is now a requirement for the fulfilment of various criteria. A prominent difference between TRAC and DRAMBORA is that the former is well placed for under-taking external audits (with a view to certification), while the latter concentrates more on self-assessment. Consequently, the processes and procedures tend to vary considerably.

2.4 Recommendations

- 1) At the current stage of development of the eCrystals data repository we recommend self-assessment using the DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) toolkit as an instrument. The audit process in many ways is more important than actual certification, since it allows repositories to analyse and respond to their archives' strengths and weaknesses in a systematic fashion. Also, DRAMBORA takes a more quantified approach to assessing repositories and would therefore work best for an established repository looking for self-assessment. TRAC (Trustworthy Repositories Audit and Certification Checklist), on the other hand is more open-ended and exploratory, taking into account vision and goals and plans for a repository and therefore more suited to repositories with an established long-term archival and preservation mandate.
- 2) Due to the recent rapid developments in this area, as well as the estimated time, effort and cost of undertaking an audit (the DRAMBORA documentation estimates 28-40 hours, depending on the scope and objectives), we have been unable to complete an audit by the end of Phase 3 (June 2007). In addition, there is a newer and hopefully more lightweight version of the toolkit due out in Sept. 2007, including an online tool. Our primary recommendation is therefore to engage the DCC (Digital Curation Centre) audit and certification team with regard to the audit of the eCrystals repository once the newer release is published in Sept. 2007; they will also have built up experience of applying DRAMBORA during several pilot repository audits by that time.
- 3) It is clear that the eCrystals data repository will require formulation of long-term commitments and objectives with regard to deposit agreements as well as expected services. However, it is recognised that making policy commitments is difficult in an academic environment, which operates under a régime of short-term contracts and funding cycles. In addition, it is worth bearing in mind that formal commitments may well entail legal liabilities. In this respect it is important to secure adequate backing from the host institution, in this case the University of Southampton. Perhaps the repository should look towards providing a low-cost subscription based service to its designated community. An alternative is to appeal to and negotiate with the IUCr [6], and other prominent organisations, with the aiming of gaining additional support and working towards a community supported sustainability plan for crystallography data.

- 4) Having examined the criteria being used in the various audit checklists, it is clear that there is a need to establish the scope and objectives of an audit more explicitly and to relate them to eCrystals more closely, so that the greatest benefit can be gained from the process.
- 5) For a long-term repository it would be beneficial to have regular audits, which verify periodically the proper functioning of records management procedures and systems and the authenticity and reliability of the records kept. Such monitoring is also useful in building up a profile of the repository over time in the face of a continuously changing environment. We suggest that a self-audit be under-taken at a frequency of once a year to enable the repository to keep abreast of developments in community standards and make sure that the technological infrastructure conforms to widely adopted standards.
- 6) In order to gain the trust of its designated community, it will be necessary for the eCrystals data repository to demonstrate compliance with the audit criteria through documentation (evidence), transparency (open examination of the evidence), adequacy (degree to which the evidence meets the vision and goals) and measurability.
- 7) Sustainability issues will be paramount in the development of the eCrystals Federation. It will be worth considering a model such as that of LOCKSS (Lots of Copies Keeps Stuff Safe) and/or CLOCKSS (Controlled LOCKSS) [30] in order to engage the crystallography community in the preservation of its valuable data.
- 8) It will also be necessary to further analyse the results of the audit to determine actions with regard to future development of the eCrystals data repository in the context of a federation

3. The Open Archival Information System (OAIS)

The development of the *Reference Model for an Open Archival Information System* (OAIS) has been led by the Consultative Committee for Space Data Systems (CCSDS). It was adopted as an ISO standard (ISO 14721:2003 [14]) in 2003. The word "Open" in the title refers to the mechanism used in the development of the model (i.e. in an open forum) rather than to the open availability of the content in an OAIS, so the model is equally applicable to dark as well as open archives. The model has recently undergone an "open" review process and a revision is imminent.

The Reference Model establishes a common framework of terms and concepts for use in the preservation of information and is not intended as a blueprint for implementation purposes. It identifies the environment within which an OAIS operates as well as its basic functions. The standard also defines a functional model as well as an information model and information flow model. Within the preservation community, it has established itself as an important standard, influencing: the development of preservation metadata; architectures and systems design of repositories; and conformance criteria for archival repositories.

The OAIS standard covers a wide range of issues relating to the operating environment of an archive or repository. For example it identifies several mandatory responsibilities:

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure its long-term preservation
- Determining the "designated community" (see section 3.1)
- Ensuring that information is "independently understandable"
- Following documented policies and procedures
- Making the preserved information available (dissemination)

In addition, it recommends the setting up of conformance and certification processes and has been built upon by several groups working in this area (notably the RLG-NARA and NESTOR work, see section 1)

The standard also contains information relating to:

- Archival Information Units and Archival Information Collections
- Information Package transformations, e.g. for ingest and access
- Preservation perspectives:
 - Migration such as refreshment, replication, repackaging and transformation
 - Preservation of the look and feel (e.g. emulation, virtual machines)
- Archive interoperability for example in P2P applications or federations

However, our main concern in this study relates to the concept of “representation information” and the OAIS functional and information models.

3.1 Important Concepts

OAIS: an archive, consisting of an organization of people and systems, which has accepted the responsibility to preserve information and make it available for a *designated community*.

Designated Community: a set of stakeholders and users served by the OAIS. In particular, a group of potential consumers who are capable of understanding a particular set of information. The designated community may be composed of multiple user communities and is subject to change over time.

Knowledge Base: a set of information, incorporated by a user or system, which allows that user or system to understand the received information; this is also likely to vary over time.

Information Object: results from *representation information* being applied to a data object (see section 3.4 below). It is important to appreciate that the OAIS model is concerned with preserving both the meaning and reusability of an information object.

Representation Information (RI): this is a very broad concept, encompassing *any* information required to render, interpret and understand (in our case, digital) data. For example, it may be a technical specification, or a data dictionary or a software tool.

Information Package: within an OAIS, information is encapsulated in packages comprising: content information, preservation description information and *packaging information*.

Packaging Information: this type of information comprises data relating to one of the processes: submission (SIP); archival (AIP) or dissemination (DIP) (see Figure 1).

Preservation Description Information: metadata deemed of particular significance to preservation issues and in particular *reference, provenance, context and fixity information* (see section 4.1).

3.2 Functional Model

The OAIS standard has a functional model comprising several entities within the operating environment of an OAIS (shown in purple in Figure 1):

Ingest: services and functions that accept SIPs from Producers; prepare AIPs for storage, and ensure that AIPs and their supporting Descriptive Information become established within the OAIS.

Archival Storage: services and functions used for the storage and retrieval of AIPs.

Data Management: services and functions for populating, maintaining, and accessing a wide variety of information.

Administration: services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis.

Preservation Planning: services and functions for monitoring the OAIS environment and ensuring that content remain accessible to the designated community.

Access: services and functions, which make the archival information holdings and related services visible to consumers.

The model places much importance on the function of *Preservation Planning*, which is the reason for its predominance in the digital preservation domain.

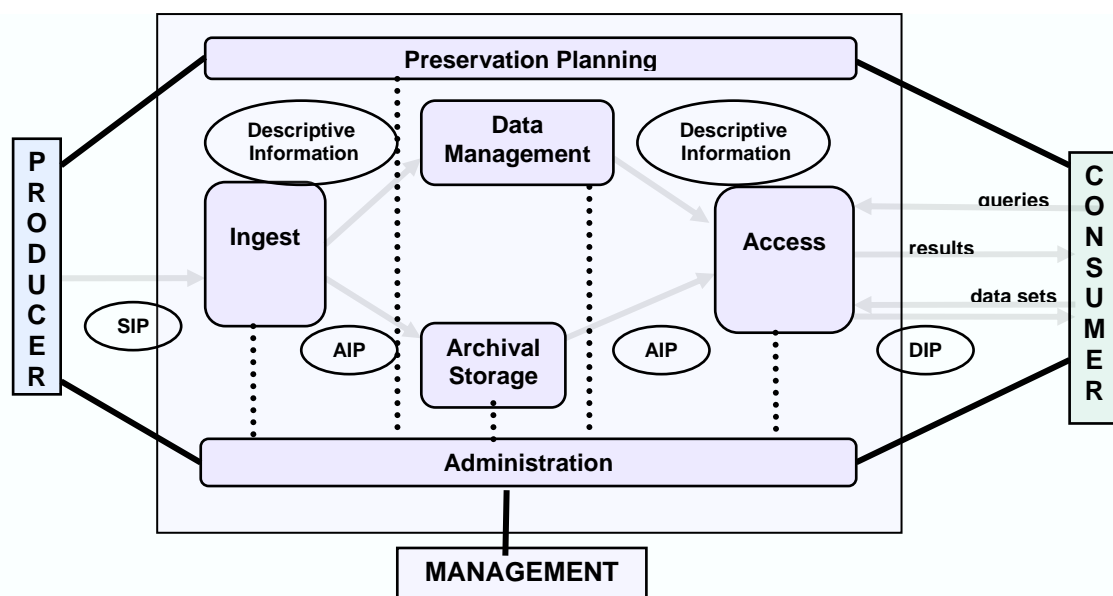


Figure 1: OAIS Functional Entities, reproduced from Figure 4-1 in the OAIS Reference Model [14]

3.4 Information Model

Information in the Reference Model is regarded as being a combination of Data and RI. The UML diagram in Figure 2 illustrates this concept. An Information Object is composed of a Data Object that is either physical or digital, as well as the RI that allows for the full interpretation of the data into meaningful information. Furthermore, any piece of RI may also be a digital object which itself needs its own RI, thus creating a Representation Information Network.

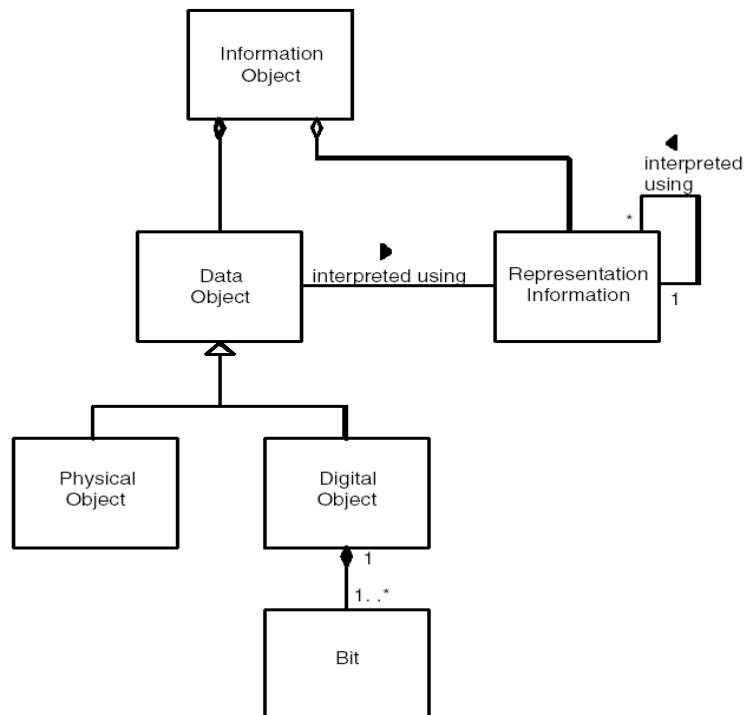


Figure 2: An OAIS Information Object, reproduced from Figure 4-10 in the OAIS Reference Model [14]

3.5 Representation Information and Networks

RI is defined as essentially whatever is required to allow a digital object to be converted to an information object. As explained in section 3.1, RI can comprise *any* information required to render, interpret and understand digital data, including: file formats, software, algorithms, standards, semantic information etc. It is essential that RI itself is curated and preserved to maintain access to other digital data.

RI is recursive in nature enabling Representation Information Networks to be built up. It is expected that the recursion will terminate for a particular designated community when the RI can be understood using that designated community's knowledge base. A problem with RI is that the amount needed for a particular object could be vast and impractical to collect in reality. It is for this reason that the concept of the designated community is so important; it enables a limit to be placed on the amount of RI which it is necessary to capture at any particular time.

In Figure 3 we see that the OAIS Model identifies three main types of RI: structural, semantic and other. Structure information describes the composition of the Data Object whilst semantic information adds meaning to specific elements. All other RI is classified under the "other" category and includes software, algorithms, standards, time varying information or actions and processes, etc.

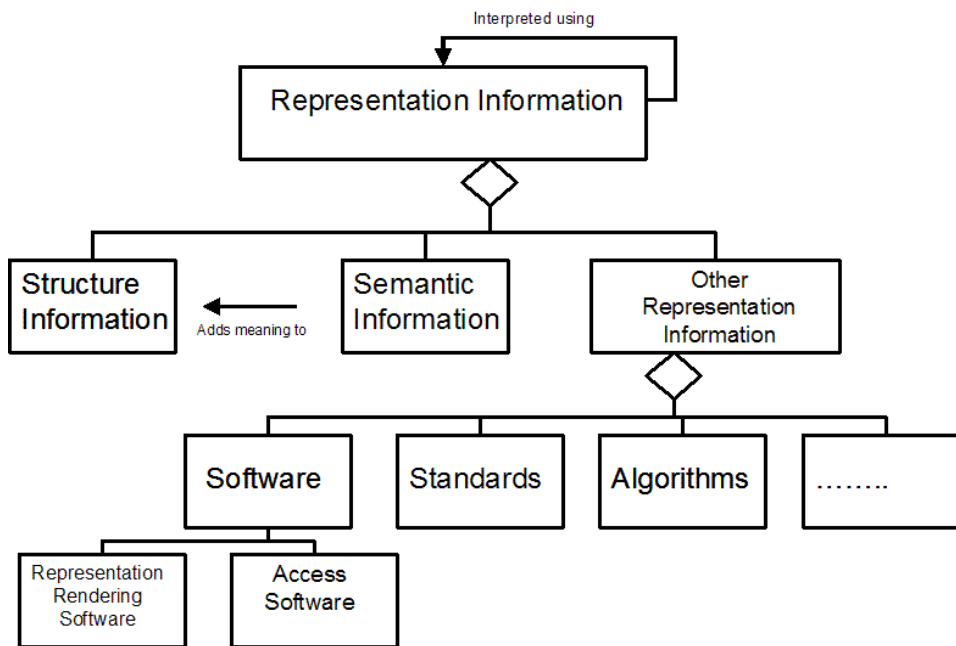


Figure 3: Types of Representation Information in the OAIS Reference Model [14]

3.7 The DCC Registry/Repository of Representation Information (RRoRI)

The DCC is developing a registry/repository of RI (RRoRI). This is not intended to be a data repository as such, but an authoritative source of RI for the community responsible for the collection, curation and management of data. The primary function is to provide and share information that enables managers of digital information to make informed decisions with regards to curation strategies. The RRoRI aims to make relevant RI available in a readily accessible manner to third parties.

This prototype registry/repository is currently being implemented as a proof-of-concept demonstrator [31]. RRoRI is heavily based on the ideas in the OAIS Reference Model and is detailed in the document, *DCC Approach to Digital Curation* [32]. The work centres on the notion that RI is the key to long-term access to digital information and that the RRoRI should be OAIS compliant.

The collection and maintenance of suitable RI mitigates the difficulties related to the preservation of understandable information. Data formats, software, standards and programming languages become obsolete; the documentation for these is often poor or non-existent; and the specialised knowledge needed to manipulate the data often disappears with time.

In order to help avoid duplication, share resources, coordinate access and minimise effort, a dedicated, well engineered and distributed network of RI is needed. However, submitters of data to an archive must be able to identify the RI needed by the end users in order to be able to work with the data. Much emphasis is placed on interoperability and automated use, the vision being to have a global, distributed network of RI which provides an infrastructure of reliable and trusted RI which other archives can rely on and use [32, 35].

To enable access to RI by third parties, the DCC-development team have cultivated the concept of an RI Label, which is used to connect RI to a particular Data Object. This label

provides a mechanism for combining individual RI components and may be a structured digital object itself (to cope with the packaging of multiple objects).

At present the prototype registry/repository [36] caters for: viewing RI already registered; registering of new RI; creation of a new RI Label; and adding a classification entry for different types of RI.

The huge burden of collecting and maintaining adequate RI requires that a global, distributed network of RI be developed; to this end the DCC-Development Team is collaborating with projects such as PRONOM [37] (developed by the National Archives) and the Global Digital Format Registry (GDFR) [38], developed by the Digital Library Federation (DLF). Both of these registries focus on provision of details about file formats (essentially the Structure type of RI).

3.7.1 Types of Representation Information

As comprehensive RI is needed to preserve access to information, it is necessary to understand the variety of forms RI may take. It will be necessary to identify what is a composite part of the data object, and what is required to enable and assist access to the information content.

Structure Information

In a digital era, structural information manifests itself largely in the form of digital file formats for text, images, audio, moving images, datasets, 3D models as well as time-varying or dynamic data. It is useful to distinguish between formats which are used mainly for rendering (for human consumption) and formats used for automated processing [32]. The former include many commercially based formats such as the succession of Microsoft Word formats; the details of such formats are likely to be proprietary and difficult or impossible to obtain. The latter are more likely to be simpler, with open source access software. Community standards in the use of structural information play an important part in establishing the knowledge base of a particular designated community.

Formal descriptions of file formats are useful in enabling automated processing, for example using the EAST [39], FLAVOR [40, 41], or DFDL [42] languages.

However, structure information may not always be available at the time of creation of digital data, and so additional effort may be required to generate it. Some digital objects, such as those created by proprietary software, can also have unknown structure. In this case, the original software, or some equivalent application, may be required to enable access.

Semantic Information

Semantic Information provides additional meaning to the contents of a digital object. For example, it may simply define the headers of a spreadsheet table, declaring that data values have been measured in a particular unit, or it may define complex relationships between objects. This category includes data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri.

Other Information

Other types of RI identified by the DCC Development Team include algorithms, software, standards, time dependent information, actions and processes. It is a characteristic of some datasets that they change over time and the state at each particular moment in time may be important (e.g. climate data or stock exchange data).

3.8 Representation Information in eCrystals

A designated community has, at any particular time, a particular knowledge base. For a specific designated community this knowledge base will evolve with time; in addition the definition of the appropriate designated community for a dataset may also change over time. The importance of identifying the designated community for a data object or more likely a collection in an archive is that it allows the archive to limit the amount of RI required for any particular digital information. The designated community for the eCrystals data repository is well defined; it is the worldwide crystallography community.

The implicit and dynamic nature of knowledge means that a knowledge base may be difficult to define; however, one possibility is to describe it as a set of familiar software applications, community standards, contextual descriptions and topic categorisations.

The eBank-UK project has constructed an institutional data repository (eCrystals) that makes available the derived and results data from a crystallographic experiment (the raw data is stored in the ATLAS data store):

“The information contained within each entry of this archive is all the fundamental and derived data resulting from a single crystal X-ray structure determination, but excluding the raw images. The results have not been externally refereed, but the information supplied should enable any reader to check the reliability and validity directly, since all the files provided are freely available for download.” eCrystals website [4].

This data repository comprises a public and a private part; through the use of an embargo schema, data can be stored as in a dark archive and reviewed periodically for conversion to open access. For the rest of this section we concentrate on the openly accessible part of eCrystals, although it should be borne in mind that RI for dark archives is as equally important for subsequent access.

The primary aim of the repository is to make available and encourage the sharing of data, which is generated throughout the experiment pipeline. The screen shot below in Figure 4, shows an example of the type of information that is stored in the repository. The top three processes (Final Result, Validation and Refinement) comprise community adopted standard file formats. In particular the CIF (Crystallographic Information File) format [43] is used within the community as an interchange format and is supported by the IUCr-International Union of Crystallographers (publisher and learned society within the domain). CIF is a publishing format as well as being structured and machine-readable; it is capable of describing the whole experiment and modelling processes (but cannot provide a reference back to the raw data file {.hkl}). Associated with the CIF format is the *checkCIF* software that is widely used within the designated community and the eCrystals data repository to validate CIF files; it is made available as an open web service by the IUCr [44].

The other type of file format included in the Final Result is a CML (Chemical Markup Language) encoding. Between the CIF and the CML file, they provide a complete description of the molecule and the chemistry of the molecule (CCDC need both in order to generate an entry in their database—the CSD). The {.cml} file is not currently produced by any software used in a crystallographic study but is generated from a {.mol} file which is an intermediate file found in the “Other Files” category. The .mol file may be generated from the {.cif} file. These file format conversions are performed according to well defined standards using the OpenBabel software obtainable from SourceForge.

The data collection, processing and solution stages are the main areas involving the work-up of the original data. The data collection stage provides JPEG files as representations of the raw data, but also proprietary formats generated by specific instrumentation that may be in

use. This stage may also have an HTML report file associated with it, providing information relating to machine calibrations and actions and how the data was processed.

The main result of the processing stage is a standardised ASCII text file {.hkl}, which has become a historical de facto standard within the designated community through its requirement by the SHELXL software. The SHELXL software produces both an output {.res} and a log file in ASCII text format. The solution stage results in a log file {.lst} comprising information relating to the computer processes that have been run on the data by the SHELXS software and a free-format ASCII text file {.prp}, which is generated by software (XPREP). There are approximately six versions of SHELXS and SHELXL, which are in use by 80-90% of the community. SHELXS and SHELXL are both commercially and openly available and currently being redeveloped.

3.8.1 Populating the RRoRI with eCrystals Data RI

This work is currently in progress and involves ingesting representation information into the RRoRI using a prototype GUI tool. It entails investigating representation information networks for crystallography data and experimentation with the use of the RRoRI ingest tool.

University of Southampton **Crystal Structure Report Archive**

Home
About
Browse
User Area
Help

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: D.B. Davies^a, R.A. Shaw^a, A. Kilic^b, M. Odlyha^a and A. Uslu^b.

Data Collection: S.J. Coles^c, L.S. Huth^c and M.E. Light^c.

Structure Determination: S.J. Coles^c, J.S. Rutherford and M.B. Hursthouse.

Birkbeck College^a
Gebze Institute of Technology^b
University of Southampton^c

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2

Compound Class: Inorganic
Keywords: cyclophosphazene, phase transition, variable temperature
Creation Date: 28 March 2007
Deposited By: Dr Simon J Coles
Deposited On: 28 March 2007

Available Files

File Name	Size
2005sjc0007.cif	11k
2005sjc0007.cml	4k
2005sjc0007_checkcif.htm	9k
2005sjc0007.res	5k
2005sjc0007_xl.lst	29k
2005sjc0007.prp	5k
2005sjc0007_xs.lst	44k
2005sjc0007.hkl	532k
2005sjc0007.htm	11k
2005sjc0007_0kl.jpg	91k
2005sjc0007_h0l.jpg	87k
2005sjc0007_hk0.jpg	79k
2005sjc0007_crystal.jpg	17k
2005sjc0007.doc	186k
2005sjc0007.fcf	138k
2005sjc0007.inchi	1k
2005sjc0007.ins	4k
2005sjc0007.mol	2k
2005sjc0007.p4p	1k
2005sjc0007_ellipsoid.gif	21k

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Refinement

Solution

Processing

Data Collection

Other Files

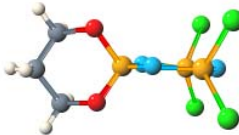


Figure 4: An example eCrystals Data Structure Report (the graphic of the molecule can be manipulated interactively)

3.9 Recommendations

- 1) eCrystals should solicit providers and developers of work-up software (SHELXS, SHELXL, XPREP etc.) to provide and maintain extensive descriptions of their file formats.
- 2) The export of raw data in the interchangeable standard format, imgCIF, by crystallographic instrumentation software is recommended.
- 3) In order to serve the crystallography community adequately it will be necessary to look at RI in the whole of the crystallography domain –not just the eCrystals repository. In particular the requirements of the CCDC, IUCr, RSC, Chemistry Central and Reciprocal Net and third party harvesters (such as Chemical Database Service) will need to be taken into account.

CCDC (Cambridge Crystallographic Data Centre) is a professional body with an international subject repository for crystal data; its Crystal Structure Database (CSD) provides federated searching across many chemistry databases. IUCr (International Union of Crystallography) is the learned society representing crystallography; it is a publisher of eight journals and maintains standards for communicating and representing crystal structures. The RSC (Royal Society of Crystallography) is also a key publisher in the field and Chemistry Central is an emerging Open Access publisher who will operate a repository to store and link data relating to publications in their journals. Reciprocal Net [45] is a distributed database used by research crystallographers to store information about molecular structures; much of the data is available to the general public.

- 4) We would advise that the eCrystals repository (and Federation) develop a formal deposit, ingest, validation and dissemination policy, to provide depositors and users with a clear indication of the service level that the eCrystals data repository and federation aims to provide.
- 5) It would also be useful to consider what type of services the eCrystals Federation would like to see developed on top of the RRoRI e.g. file-format obsolescence notification, choice of migration strategies etc.

4. eBank-UK Application Profile and Preservation Metadata

Details of the development and current status of the eBank-UK Metadata Application Profile are readily available on the project website [46]. Phases 1 and 2 of the project did not focus on curation and preservation issues; consequently, much of the necessary metadata is currently absent from the profile.

According to Lavoie and Gartner [47], preservation metadata is information that supports and documents the long-term preservation of digital materials. It addresses an archived digital object's:

Provenance: documenting the custodial history of the object

Authenticity: validating that the digital object is in fact what it purports to be, and has not been altered in an undocumented way

Preservation activity: documenting the actions taken to preserve the digital object, and any consequences of these actions that impact its look, feel, or functionality

Technical environment: describing the technical requirements, such as hardware and software, needed to render and use the digital object

Rights management: recording any binding intellectual property rights that may limit the repository's ability to preserve and disseminate the digital object over time.

The OAIS Reference Model [14] has been influential in the development of preservation metadata; it provides a high-level overview of the types of information needed to support digital preservation, including: representation information; preservation description information (reference, context, provenance and fixity information, see section 4.1); packaging information and descriptive information. These types of information can be considered as general categories of metadata, which are required to support the long-term preservation and use of digital materials; they have served as the starting point for a number of preservation metadata initiatives.

Over the years, a number of institutions and projects have investigated and developed preservation metadata element sets (e.g. National Library of Australia [48], CEDARS project [49], NEDLIB Project [50]). However, in 2002, the OCLC/RLG Preservation Metadata Framework Working Group consolidated existing expertise in the form of a preservation metadata framework [51]. Using the broad categories of information specified in OAIS as a starting point, the Framework enumerated the types of information falling within the scope of preservation metadata. Release of the Framework prompted interest in a more practical and implementation-oriented way forward.

In June 2003, OCLC and RLG therefore sponsored a second working group: PREMIS (PREservation Metadata: Implementation Strategies). The membership included more than thirty international experts in preservation metadata. The remit of the group was to:

- Define a core set of implementable, broadly applicable preservation metadata elements, supported by a data dictionary
- Identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata in digital archiving systems.

In September 2004, PREMIS released a survey report describing current practice and emerging trends associated with the management and use of preservation metadata to support repository functions and policies [52]. In May 2005, PREMIS followed up the survey report with the *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group* [53]. The report includes: the PREMIS Data Dictionary v1.0; an accompanying report, which provides context and an underlying data model; usage examples and a set of XML schema to support use of the Data Dictionary. In addition, a maintenance activity has been set up to manage the evolution of the Data Dictionary [54].

More recently (June 2007), a report on *Implementing the PREMIS data dictionary: a survey of approaches* by Deborah Woodyard-Robinson [55] examines the take-up of the PREMIS Data Dictionary and the implementation issues that have been encountered by 16 repositories.

4.1 OAIS Preservation Description Information

OAIS defines Preservation Description Information as comprising several different types of information:

Reference: One or more mechanisms used to provide assigned identifiers for unambiguous access to content. Examples include: object identifier; a journal reference; a bibliographic description or a persistent identifier.

Provenance: Documents the history of the content information including: any changes that may have taken place since it was submitted and who has had custody of it. It provides users some assurance as to the likely reliability of the content information.

Context: Documents the relationships of the content information to its environment and other content information. Examples include: calibration history; relationship to other data sets; pointers to related documents etc.

Fixity: Provides data integrity checks including validation/verification keys used to ensure that the particular content information object has not been altered in an undocumented manner. Examples include: special encoding and error detection schemes that are specific to instances of the content object (e.g. checksums).

4.2 PREMIS Data Dictionary

Although the OAIS Reference Model and its concept of preservation description information (PDI) remains the conceptual foundation for the PREMIS data dictionary, the data model does in fact diverge and is instead derived from the work of the National Library of New Zealand on preservation metadata [58]. The PREMIS data dictionary provides an intermediate stage in between OAIS PDI and the actual application specific implementation; its major functions are to cater for data exchange and interoperability.

In establishing a “core” set of preservation metadata, the PREMIS group uses a practical definition i.e. “things that most working preservation repositories are likely to need to know in order to support digital preservation”.

The dictionary is implementation independent and uses an entity-relationship data model based on five types of entities to provide semantic information. In broad terms, *Entities* are involved in digital preservation activities and consist of: intellectual entities, objects, rights, agents and events. *Relationships* are statements of association between instances of entities; the direction of the arrows shows the direction of the relationship (double-headed arrows indicate reciprocal links). *Semantic Units* are used rather than “metadata elements” in order to emphasise the “need to know” aspect; they are properties of an entity and therefore have values. *Representation* is the set of files required for an intellectual entity to be displayed, played or otherwise made usable to a human.

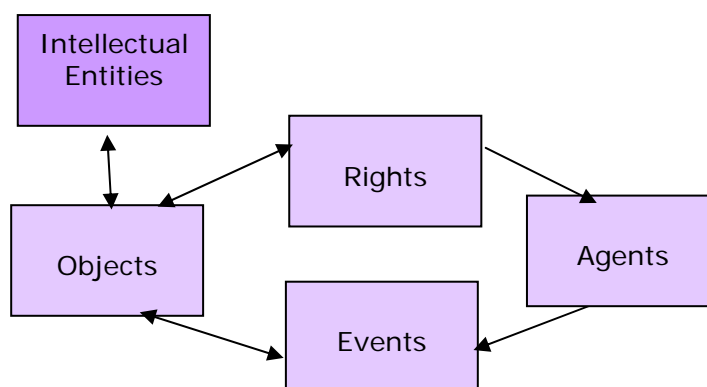


Figure 5: PREMIS Data Dictionary Data Model [53]

An Object: is a discrete unit of information in digital form. Objects are the resources that the repository preserves and may be one of the following types:

FILE: a named and ordered sequence of bytes that is known by an operating system.

REPRESENTATION: the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.

BITSTREAM: contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes.

An Intellectual Entity: is a coherent set of content that is reasonably described as a unit, for example: a particular book, map, photograph, or database. It may include other Intellectual Entities (e.g. as a website includes a web page). It could also have one or more digital representations.

An Event: is an action that involves at least one object or agent known to the preservation repository. It is concerned with the details necessary to document digital provenance

An Agent: can be a person, organization, or software program associated with preservation events in the life of an Object. Agents influence an Object indirectly through an Event. Agents are not defined in detail in PREMIS since they are not considered core preservation metadata beyond that required for identification.

Rights: comprise an agreement with a rights holder that allows a repository to take action(s) related to Objects in the repository. The data dictionary does not specify a full rights expression language, however basic statements such as Agent A grants Permission P for Object B are catered for. Note that PREMIS deals only with rights and permissions related to preservation activities, leaving aside those concerned with access and dissemination.

The Data Dictionary offers Semantic Units for Objects, Events, Agents and Rights. Intellectual entities are considered to be well served by other descriptive metadata. Examples of Semantic Units related to objects include:

objectCategory (mandatory)

Values: representation, file, bitstream

preservationLevel

What preservation treatment/strategy the repository plans for this object

Could be a business rule only relevant in a given repository

Examples: full, bit-level

Examples of Semantic Units relating to object creation information:

creatingApplication

Information about an application that created the object

Container with 3 subunits: name, version and date

Applies to objects created externally or by a repository

Repeatable if more than one application processed the object

Example: MS Word 2000 [date created]

originalName

Name of object as submitted to or harvested by repository

Supplements repository supplied names

Only applicable to files

Example: sip/book/N419.pdf

4.3 Preservation Metadata in eCrystals

Whilst the PREMIS Data Dictionary remains the authoritative reference for the development of preservation related metadata and hence interoperability, it is also useful to learn from and build on the experience of various projects and initiatives that have already attempted to create such metadata. In particular, several review and survey documents have recently become available, including: *Implementing the PREMIS data dictionary: a survey of approaches*, A Report for the PREMIS Maintenance Activity [55]; *Preservation Metadata for Institutional Repositories: applying PREMIS* [63] and *Review of metadata standards in use by SHERPA DP repositories* [64]. In addition, Priscilla Caplan (co-chair of the PREMIS working group), provides a helpful overview of preservation metadata in her instalment on *Preservation Metadata*, for the DCC Digital Curation Manual [65].

It should be borne in mind that differing preservation strategies are likely to demand that distinct types of information be recorded. For example, a preservation plan based on migration will require differing information to that of one based on emulation. Hence, the preservation planning and policies of a particular repository will heavily influence the specific metadata that is to be recorded.

Nevertheless, it is possible to identify certain types of core information which is currently regarded as “things that most working preservation repositories are likely to need to know in order to support digital preservation” [53]. This necessarily cuts across the categories which are typically used to identify different types of metadata (administrative, descriptive, structural and technical). According to Caplan [65], such a core set of preservation metadata should include the following:

File Format identification: it is crucial to record information relating to the format of a digital file. Since file extensions and MIME types do not provide sufficient granularity or distinguish between versions it will be necessary to use file format registries such as PRONOM [37] or the GDFR [38]. For automated extraction of format information, tools such as JHOVE [66] and DROID [67] can be used. There is also a need to take account of the standards and formats within the knowledge base of the designated community.

Significant Properties: these are characteristics which should be retained through future preservation activities. eCrystals will need to determine such characteristics together with depositors and its designated community.

Environment for use: environment information comprises a record of the hardware, software and any other information required to render or use the digital data. Much of this information can be associated with the file format and therefore shared between data-sets.

Fixity information: this is essential in verifying the authenticity of a file and is commonly implemented using a checksum. However, even within a single computer system, error-free transfers of data cannot be taken for granted. Depending on the policies of the eCrystals data repository and federation, it may be necessary to consider using additional checks or a more sophisticated authentication system.

Technical information: while file format and environment information encompass much of this type of metadata, there may be other technical information that may be relevant for crystallography data. For example, bit depth is important with regard to audio and image data.

Provenance: the origin and chain of custody of a digital object are important factors in the trust that users place in it; such information includes: creation information (including creator and date/time); owners; rights holders; record of actions (events and processes performed on

the object). Note that the PREMIS data model does not allow the modification of a digital object; rather the act of changing an object creates an entirely new object which is related to the source object by derivation.

It will be necessary to analyse the data model of the eCrystals data repository to establish relationships with the data model in the PREMIS Data Dictionary. However, a cursory evaluation of the PREMIS Data Dictionary suggests that the following top-level Semantic Units are likely to be of importance for the eCrystals data repository:

Object Entity:

objectIdentifier, preservationLevel, objectCharacteristics, creatingApplication, storage, environment, relationship, linkingEventIdentifier, linkingIntellectualEntityIdentifier, linkingPermissionStatementIdentifier

Event Entity

eventIdentifier, eventType, eventDateTime, eventDetail, eventOutcomeInformation, linkingAgentIdentifier, linkingObjectIdentifier

Agent Entity

agentIdentifier

Rights Entity

permissionStatementIdentifier, linkingObject, permissionGranted, grantingAgent

4.4 Recommendations

- 1) The eCrystals data repository needs to develop a preservation plan and strategy appropriate to the types of content and data that it is charged with curating.
- 2) We recommend that additional (preservation) metadata be recorded in order to take into account the Semantic Units expressed in the PREMIS Data Dictionary where relevant to the preservation plan adopted (tentative Semantic Units are suggested above, in section 4.3). This could be achieved either by revision or extension of the current eBank-UK Metadata Application Profile, or by implementing a separate eBank Preservation Metadata Profile; the appropriateness of these strategies need to be assessed, in particular in the context of the Federation.
- 3) Reference information: the eCrystals data repository currently uses Digital Object Identifiers [56] as a form of reference identifier as well as the IUPAC International Chemical Identifier (InChi) [57] as a domain identifier; it may be beneficial to analyse how DOIs will fit in with the operation of the proposed Federation.
- 4) Provenance information: versioning is the only type of information currently stored by the ePrints.org software. eCrystals will need to consider what other types of provenance information should be recorded relevant to its selected preservation strategy.
- 5) Context information: the only type of relationship recorded in ePrints.org at present is that of versioning information, again eCrystals will need to consider what other types of context information should be recorded relevant to its adopted preservation strategy.
- 6) Fixity information: in addition to the use of the checkCIF utility, there are several, simple integrity checks performed in the 'toolbox' data file manipulation and deposit software. The adequacy of these checks needs to be investigated.

- 7) At present, content is not stored in the form of OAIS Information Packages, as such; although, it can be disseminated in the form of METS packages. It will be necessary to consider whether this is adequate as far as the requirements of the Federation are concerned.
- 8) Rights Information (copyright, IPR, preservation rights): we expect that it will be necessary to revisit rights metadata in the context of a Federation.
- 9) Interoperability and data exchange within the eCrystals Federation is likely to be easier and more successful if consensus can be achieved on the use of the eBank-UK Metadata Application Profile.
- 10) Creation and maintenance of metadata is an expensive process; the eCrystals repository and Federation will need to address the question of who should create preservation metadata and investigate to what extent such data can be generated, extracted and maintained automatically.

5. ePrints.org Conformance to OAIS

From a preservation perspective it is paramount that the software platform underlying a repository be able to support the concepts in the OAIS Reference Model. Despite the lack of a formal process for certification, many repository software platforms and preservation tools claim OAIS compliance or conformance, for example: DSpace [60], OCLC Digital Archive [61], METS [59], LOCKSS [30] and FEDORA [62]. Indeed, the high-level and conceptual nature of the OAIS Reference Model does nothing to discourage such claims.

The eCrystals data repository is built using the ePrints.org software [12] (currently using version 2 with an imminent migration to version 3). The software certainly supports the OAIS concepts of ingest, archive and retrieval at a high-level, however, it does not have a concrete notion of information packages as such (although it does allow the export of information encapsulated in METS packages). Note that the OAIS Model describes information packages as a “logical” concept.

With regard to the different types of OAIS Preservation Description Information:

Reference information: ePrints.org allows use of any identifier scheme that an application may wish to use; eCrystals uses InChi and DOI.

Provenance information: ePrints.org automatically tracks changes and modifications made to the information object once it has been ingested into the repository (e.g. versioning information). The software does not automatically store other provenance data, however ePrints.org does now support a range of licenses under which content can be made available.

Fixity information: Checksums are generated on ingest into an ePrints repository. File format checking and validation are currently not performed (this is done at the application level by the eCrystals repository using checkCIF). Virus checking is due to be introduced into ePrints.org.

Context information: The only type of relationship recorded in ePrints.org at present is that of versioning information.

6. Conclusions

This scoping report is an attempt to highlight preservation and curation issues that are likely to be significant in moving the eCrystals data repository from a pilot research service to a more sustainable long-term resource in the crystallography domain. It is clear that such a commitment is multi-faceted and requires a considerable amount of additional effort and resources. Consequently, it would be useful to undertake a cost/benefit analysis of potential preservation strategies and the level of risk mitigation that each offers.

Ideally, preservation, curation and sustainability issues should be considered at the outset; at the planning and design stage of a repository. We appreciate that this not always possible when developing a prototype research repository. Indeed, it could be argued that the significant cost and effort required cannot be justified until the resource has proven its usefulness to the designated community.

We have also tried to cast light on the types of issues that are likely to be of relevance to the development of the proposed eCrystals federation. In particular, it should be noted that community and organisational support are essential for the long-term sustainability of digital data and that this will entail both advocacy and collaborative working.

Acknowledgements

We would like to thank Les Carr and Chris Gutteridge for discussions relating to the ePrints.org software and Brian McIlwrath for development of the RRORi ingest tool.

References

1. The eBank-UK Project, <http://www.ukoln.ac.uk/projects/ebank-uk/>
2. Liz Lyon, *eBank UK: Building the links between research data, scholarly communication and learning*, Ariadne, Issue 36, July 2003
<http://www.ariadne.ac.uk/issue36/lyon/>
3. The Open Archives Initiative, <http://www.openarchives.org/>
4. The Crystal Structure Report Archive –eCrystals Data Repository,
<http://ecrystals.chem.soton.ac.uk>
5. eBank-UK: Phase 3 Project Plan, to appear
6. The International Union of Crystallography (IUCr), <http://www.iucr.org/>
7. The Cambridge Crystallographic Data Centre (CCDC), <http://www.ccdc.cam.ac.uk/>
8. The Digital Curation Centre (DCC), <http://www.dcc.ac.uk>
9. Philip Lord and Alison Macdonald, *e-Science Curation Report, Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*, prepared for The JISC Committee for the Support of Research (JCSR), 2003, http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf#search=%22e-Science%20curation%20report%22
10. JISC Circular 6/03 (Revised), <http://www.dcc.ac.uk/docs/6-03Circular.pdf>
11. Core Requirements for Digital Archives,
<http://www.crl.edu/content.asp?11=13&12=58&13=162&14=92>
12. ePrints.org, EPrints for Digital Repositories, <http://www.eprints.org/>
13. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System*, ISO:14721:2002,
<http://public.ccsds.org/publications/archive/650x0b1.pdf#search=%22OAIS%20mode1%22>
14. ISO 14721:2003, *Open Archival Information Reference Model*,
<http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683>

15. *Trusted Digital Repositories: Attributes and Responsibilities*, An RLG OCLC Report, May 2002, <http://www.rlg.org/legacy/longterm/repositories.pdf>
16. A.J.G. Hey and A. E.Trefethen, *The Data Deluge: An e-Science Perspective*, in F. Berman, G.C. Fox and A.J.G. Hey Eds. *Grid Computing-Making the Global Infrastructure a Reality*, chapter 36, pages 809-824, 2003, Wiley & Sons.
17. *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, May 1, 1996, <ftp://ftp.rlg.org/pub/archtf/final-report.pdf>
18. The Deming Cycle, http://en.wikipedia.org/wiki/Shewhart_cycle
19. *An Audit Checklist for the Certification of Trusted Digital Repositories*, RLG/NARA, August 2005, <http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf>
20. Digital Preservation Europe (DPE), <http://www.digitalpreservationeurope.eu/>
21. Network of Expertise in Long-Term Storage of Digital Resource, <http://www.langzeitarchivierung.de/index.php?newlang=eng>
22. *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)*, Version 1.0, Feb. 2007, Center for Research Libraries and RLG Programs (revised and expanded version of The Audit Checklist for the Certification of Trusted Digital Repositories, originally developed by RLG-NARA Digital Repository Certification Task Force), <http://www.crl.edu/content.asp?11=13&12=58&13=162&14=91>
23. *DINI-Certificate Document and Publications Services 2007*, Draft Version, Version 2, Electronic Publishing Working Group, Sept. 2006 http://www.dini.de/documents/DINI_certificate_eng_2006-10-12_draft.pdf
24. Suzanne Dobratz and Astrid Schoger, *Digital Repository Certification: A Report from Germany*, DINI/NESTOR, October 2005 <http://edoc.huberlin.de/oa/articles/reh7CbxRopdUA/PDF/23yn183UoMBU.pdf>
25. *DRAMBORA -Digital Repository Audit Method Based on Risk Assessment*, March 2007, Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), <http://www.repositoryaudit.eu/>
26. Seamus Ross and Andrew McHugh, *The Role of Evidence in Establishing Trust in Repositories*, *Dlib* magazine 12(7/8), July/August 2006 <http://www.dlib.org/dlib/july06/ross/07ross.html>
27. S Ross and A McHugh, *Audit and Certification: Creating a Mandate for the Digital Curation Centre*, *RLG Diginews*, 9.5, ISSN 1093-5371, http://www.rlg.org/en/page.php?Page_ID=20793#article1
28. Hans Hofman, Raivo Ruusalepp and Andrew McHugh, *Context and Development of the DRAMBORA Toolkit*, DCC Tutorial, Building Trust in Digital Repositories Using the DRAMBORA Toolkit, 27 April 2007, The British Library, London
29. International Repository Audit and Certification Birds of a Feather Group Wiki, <http://wiki.digitalrepositoryauditandcertification.org/bin/view>
30. LOCKSS –Lots of Copies Keeps Stuff Safe, <http://www.lockss.org/lockss/Home>
31. The DCC Registry of Representation Information (RoRI), <http://registry.dcc.ac.uk/>
32. DCC Development Team, *DCC Approach to Digital Curation*, <http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCApproachToCuration>
33. DCC Development Team, *DCC Label Report*, <http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCInfoLabelReport>
34. David Giarretta, Manjula Patel, Adam Rusbridge, Stephen Rankin, Brian McIlwrath, *Supporting e-Research Using Representation Information*, Proceedings UK e-Science All Hands Meeting, 2005, <http://www.allhands.org.uk/2005/proceedings/papers/447.pdf>
35. David Giarretta, Stephen Rankin, Brian McIlwrath, Adam Rusbridge, Manjula Patel, *Representation Information for Interoperability Now and with the Future*, Proceedings MSST 2005, pp54-58, International IEEE Symposium on Mass Storage and Systems, 20-24th June 2005, Sardinia, Italy, <http://www.soe.ucsc.edu/~elm/msst05/MSST05-Sardinia-Proceedings-2.pdf>

36. DCC Development Team, *RoRI High-Level Design*, <http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCRegRepOverall>
37. The PRONOM registry, The National Archives, <http://www.nationalarchives.gov.uk/pronom/>
38. The Global Digital Format Registry (GDFR), Digital Library Federation, <http://hul.harvard.edu/gdfr/>
39. The EAST Data Description Language, <http://east.cnes.fr/english/index.html>
40. Alexandros Eleftheriadis, *Flavor: a language for media representation*, Proceedings of the Fifth ACM International Conference on Multimedia, Seattle, Washington, US, 1997, <http://portal.acm.org/citation.cfm?id=266319>
41. Alexandros Eleftheriadis and Danny Hong, *Flavor: A Formal Language for Audio-Visual Object Representation*, Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, US, pages 816-819, 2004, <http://portal.acm.org/citation.cfm?id=1027717>
42. Data Format Description Language (DFDL), DFDL Working Group, Open Grid Forum, <http://forge.gridforum.org/projects/dfdl-wg>
43. CIF -The Crystallographic Information File, <http://www.iucr.org/iucr-top/cif/>
44. IUCr checkCIF validation service, <http://checkcif.iucr.org/>
45. The Reciprocal Net, <http://www.reciprocalnet.org/index.html>
46. eBank-UK: Metadata Schemas, <http://www.ukoln.ac.uk/projects/ebank-uk/schemas/>
47. Brian Lavoie and Richard Gartner, *Preservation Metadata*, Digital Preservation Coalition Technology Watch Report, September 2005, <http://www.dpconline.org/docs/reports/dpctw05-01.pdf>
48. National Library of Australia (1999) Preservation Metadata for Digital Collections, October 1999, <http://www.nla.gov.au/preserve/pmeta.html>
49. Cedars Project: *Guide to Preservation Metadata*, <http://www.leeds.ac.uk/cedars/guideto/metadata/>
50. Catherine Lupovici and Julien Masanès, *Metadata for Long-Term Preservation*, NEDLIB project, July 2000, <http://nedlib.kb.nl/results/D4.2/D4.2.htm>
51. OCLC/RLG Working Group on Preservation Metadata, *A Metadata Framework to Support the Preservation of Digital Objects*, June 2002, http://www.rlg.org/en/pdfs/pm_framework.pdf
52. *Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community*, PREMIS Working Group, September 2004, <http://www.oclc.org/research/projects/pmwg/surveyreport.pdf>
53. *Data Dictionary for Preservation Metadata*, Final report of the PREMIS Working Group, May 2005, <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>
54. PREMIS Preservation Metadata: Maintenance Activity, <http://www.loc.gov/standards/premis/>
55. Deborah Woodyard-Robinson, *Implementing the PREMIS data dictionary: a survey of approaches*, A Report for the PREMIS Maintenance Activity, <http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf>
56. Digital Object Identifier System, DOI Foundation, <http://www.doi.org/>
57. IUPAC International Chemical Identifier (InChi), <http://www.iupac.org/inchi/>
58. *Metadata Standards Framework –Preservation Metadata (revised)*, National Library of New Zealand, June 2003, <http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised>
59. Metadata Encoding and Transmission Standard (METS), <http://www.loc.gov/standards/mets/>
60. DSpace, <http://www.dspace.org/>
61. OCLC Digital Archive, <http://www.oclc.org/digitalarchive/about/default.htm>
62. FEDORA –Flexible Extensible Digital Object Repository Architecture, <http://www.fedora.info/documents/WhitePaper/FedoraWhitePaper.pdf>

63. Steve Hitchcock, Tim Brody, Jessie M.N Hey and Leslie Carr, *Preservation Metadata for Institutional Repositories: applying PREMIS*, January 2007, <http://preserv.eprints.org/papers/presmeta/presmeta-paper.html>
64. Gareth Knight and Kirti Bodhimage, *Review of metadata standards in use by SHERPA DP repositories*, February 2006, http://www.sherpadp.org.uk/documents/wp41-metadata_standards.pdf
65. Priscilla Caplan, *Instalment on "Preservation Metadata"*, DCC Curation Manual, July 2006, <http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/>
66. JHOVE -JSTOR/Harvard Object Validation Environment, <http://hul.harvard.edu/jhove/>
67. DROID –Digital Record Object Identification, The National Archives, <http://droid.sourceforge.net/wiki/index.php/Introduction>

Appendix I –TRAC Audit Checklist

The *TRAC* checklist contains 84 criteria broken out into three main sections: Organizational infrastructure; Digital object management; and Technologies, technical infrastructure, and security. Within each of these sections are various subsections and under the subsections are the criteria themselves.

- A. Organizational infrastructure
 1. Governance & organizational viability
 2. Organizational structure & staffing
 3. Procedural accountability & policy framework
 4. Financial sustainability
 5. Contracts, licenses, & liabilities
- B. Digital object management
 1. Ingest: acquisition of content
 2. Ingest: creation of the archivable package
 3. Preservation planning
 4. Archival storage & preservation/maintenance of AIPs
 5. Information management
 6. Access management
- C. Technologies, technical infrastructure, and security
 1. System infrastructure
 2. Appropriate technologies
 3. Security

Some sample criteria are:

- A1.1 Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information.
- A2.2 Repository has the appropriate number of staff to support all functions and services.
- A3.5 Repository has policies and procedures to ensure that feedback from producers and users is sought and addressed over time.
- A5.4 Repository tracks and manages intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.
- B2.5 Repository has and uses a naming convention that generates visible, persistent, unique identifiers for all archived objects (i.e., AIPs).
- B2.9 Repository acquires preservation metadata (i.e., PDI) for its associated Content Information.
- B3.4 Repository can provide evidence of the effectiveness of its preservation planning.
- B4.4 Repository actively monitors integrity of archival objects (i.e., AIPs).
- B5.3 Repository can demonstrate that referential integrity is created between all archived objects (i.e., AIPs) and associated descriptive information.
- C1.1 Repository functions on well-supported operating systems and other core infrastructural software.
- C1.5 Repository has effective mechanisms to detect bit corruption or loss.
- C1.7 Repository has defined processes for storage media and/or hardware change (e.g., refreshing, migration).
- C1.9 Repository has a process for testing the effect of critical changes to the system.
- C3.3 Repository staff have delineated roles, responsibilities, and authorizations related to implementing changes within the system.

Appendix II–DRAMBORA Checklist

DRAMBORA is a more methodical approach to assessing the trustworthiness of a repository. A systematic process guides the auditor to identify risks to long-term preservation of repository content, and then scores each risk as a product between the likelihood of the risk occurring with the impact associated with that event. Mitigation of the risks could then be prioritized based on a descending order of the score.

The process has six stages, some with multiple tasks:

1. Identify organizational context
 - Specify mandate of your repository or the organization in which it is embedded
 - List goals and objectives of your repository
2. Document policy and regulatory framework
 - List your repository's strategic planning documents
 - List the legal, regulatory, and contractual frameworks or agreements to which your repository is subject
 - List the voluntary codes to which your repository has agreed to adhere
 - List any other documents and principles with which your repository complies
3. Identify activities, assets and their owners
 - Identify your repository's activities, assets and their owners
4. Identify risks
 - Identify risks associated with activities and assets of your repository
5. Assess risks
 - Assess the identified risks
6. Manage risks
 - Manage the risks identified

The report includes a catalog of risks taken from other checklists and repository audits that can be used to spur the thinking of the auditor.