Preservation Planning for Crystallography Data

JISC eCrystals Federation Project

WP4: Repositories, Preservation and Sustainability

Document Details

Author:	Manjula Patel (UKOLN & DCC)
Date:	25th June 2009
Version:	0.8
File Name:	eCrystals-WP4-PP-090625.doc
Notes:	Final

Executive Summary

The aim of the eCrystals Federation project is to enhance the management of crystallography data at the institution level, incorporating data generated in departments, laboratories and by individual researchers or practitioners. The project is attempting to set up a federation of institutional repositories for the management and dissemination of derived and results data from crystallographic experiments. WP4 of the project is concerned with the development of approaches to the preservation and curation of crystallography data in open repositories. We consider that preservation planning and its associated activities should be viewed as an integral part of sound data management practice.

However, one of the obstacles to proactively undertaking preservation and curation activities identified by the interim report of the Blue Ribbon Task force on Sustainable Digital Preservation and Access[1] is a fear that digital access and preservation is too big a problem to take on. Given their wide-ranging influences and multi-faceted nature, it is no surprise that curation and preservation are considered by many to be daunting tasks. Stewardship of research data is undoubtedly a substantial commitment to sign up to with major implications for funding and resources. Nevertheless, today's repository managers are under increasing pressure to provide expertise in IT skills, domain knowledge and preservation issues. We consider that expertise in all three areas is essential for the effective management of digital research data and in this report focus particularly on curation and preservation issues.

Over the last two decades a huge amount of research has been undertaken with respect to particular aspects of digital curation and preservation resulting in an accumulation of tools, frameworks and guidance; our aim is to raise awareness of these and guide repository managers to a subset which is relevant to the long-term management of crystallography data. Our objective has been to break down the process of preservation planning into manageable components, thus providing a starting point for managers of repositories to consider the likely preservation issues specific to their particular repository. We illustrate various aspects of preservation planning with reference to an exemplar crystallography data repository supported and run by the EPSRC National Crystallography Service (NCS) based at the University of Southampton.

Social, political, cultural, organizational, financial, legal and technical issues will all impact on a commitment to the long-term management of digital data. As a result preservation planning has to be undertaken in the organisational context and specific circumstances of a repository. Consequently, the solutions forged will be specific to each repository – there is no "one size fits all" solution. In addition, we recognise that there is considerable diversity in crystallography laboratory practice which will require customised preservation planning.

Nevertheless, we believe that there are generic methodologies and tools which can be applied in order to help achieve preservation plans appropriate to individual repositories. We have therefore attempted to break down the process of preservation planning into various practical aspects, including (analysis of data and workflows; evaluation of preservation requirements; defining a preservation policy; formulating a preservation strategy; recording preservation metadata; modelling costs; planning for sustainability; and regular evaluation or self-assessment). Given the complexity of the topics and issues involved (even for a sub-discipline such as crystallography) the coverage focuses on breadth rather than depth.

---

[1] *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*, Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, December 2008, http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Acknowledgements

The eCrystals Federation Project and the Digital Curation Centre (DCC) are both funded by the UK's Joint Information Systems Committee (JISC).

Revision History

Date	Contributor	Contribution
4th February 2009	M Patel	Setting up and outline of report and associated work (focus changed from DRAMBORA assessment to preservation planning)
10th February 2009	M Patel	Refinement of topics to be covered
1st-23rd April 2009	M Patel	Review of various preservation planning tools and documents
27-30th April 2009	M Patel	Restructuring and writing up of sections 1-4
11-31st May 2009	M Patel	Writing up of section 5-6.3
1st June 2009	M Patel	Writing up of sections 6.4-6.6
2nd June 2009	M Patel	Sections 6.7, 6.8; Conclusions
3rd June 2009	M Patel	Exec summary; References Circulation of first draft for comments
25th June 2009	M Patel	Took into account comments from Simon Coles Final version made available for circulation

Contents

1. Introduction....6
2. Crystallography Data....7
3. An Exemplar Repository: eCrystals@Soton....8
4. Trust and Data Stewardship....8
5. Over-arching Tools & Frameworks....10
5.1 DCC Curation Lifecycle Model....10
5.2 OAIS and other Standards....11
5.3 Audit and Certification Instruments....12
5.3.1 TRAC Checklist....12
5.3.2 DRAMBORA Risk Assessment....13
5.3.3 PLATTER Toolkit....14
5.3.4 Data Seal of Approval....15
6. Components of Preservation Planning....15
6.1 Analyse Data & Associated Workflows....16
6.2 Evaluate Preservation Requirements....18
6.3 Define a Preservation Policy....18
6.4 Formulate a Preservation Strategy....19
6.5 Record Preservation Metadata....21
6.6 Model Costs....23
6.7 Plan for Sustainability....24
6.8 Regular Evaluation and/or Self-Assessment....25
7. Conclusions....25
References....26

1. Introduction

A compelling case for the curation and preservation of digital scientific data has been built up over recent years [1, 2]. Many of these studies and reports call for the development of national and international infrastructure and support services to enable the management of data resulting from the scholarly research process [3, 4, 5].

A number of reasons can be identified for maintaining and providing ready access to research data for reuse. Data is evidential in supporting research and scholarship, providing for the verification and validation of results. Furthermore, research outputs feed into and contribute to the scholarly knowledge lifecycle based on continuous use and reuse of data [6]. In addition, well managed and curated data has the potential to be re-purposed and generate new science. Recapturing and reproducing some types of data is often difficult or even impossible, for example observational and environmental data is often unique and temporal in nature; other types of data may be cheaper to maintain than to regenerate. Some types of data have legal obligations associated with them and must be retained for certain periods of time for compliance. Furthermore, research funding bodies are becoming increasingly aware of the need to protect and enhance their investments in research by ensuring that data is made widely available so that the greatest value can be extracted from it, maximising the opportunity for reuse, cross-reference and dataset integration. They would also like to ensure that valuable datasets are stored securely and remain readily accessible to future researchers.

Global cooperation in managing research data is also becoming apparent; for example, a group of Europe's leading research libraries and technical information providers have recently established a partnership to improve access to research data on the internet [7]. The goal of this cooperation is to establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects.

The term digital curation includes the active management of digital data and research results over their entire scholarly and scientific life-time, both for current and future use. It also encompasses the notion of adding value to a trusted body of digital information as well as its reuse in the derivation of new information and the validation and reproducibility of scientific results [8]. Curation, in the first instance requires a commitment to undertake duties of stewardship. However it should be noted that such a commitment is influenced by a complex array of factors including social, political, organizational, financial and legal as well as technical issues – as a consequence, there is no "one size fits all" solution.

Subject-based and institutional repositories are emerging as the preferred means of safe-guarding the future availability and accessibility of digital information. The policies and commitments relating to the stewardship of data will inevitably vary depending on the size, type and status of the repository that is used to manage the data.

The eCrystals Federation project is concerned with setting up a federation of institutional repositories for the management and dissemination of derived and results data from crystallographic experiments [9]. WP4 of the project is concerned with the development of approaches to the preservation and curation of crystallography data in open repositories. We consider that preservation planning and its associated activities should be viewed as an integral part of sound data management practice. Repository managers are under increasing pressure to provide expertise in IT skills, domain knowledge as well as preservation issues. The objective of this report is to provide a starting point, for managers of repositories tasked

with managing crystallography data, to consider the likely preservation issues specific to their particular repository. Over the last two decades a huge amount of research has been undertaken with respect to particular aspects of digital curation and preservation resulting in an accumulation of tools, frameworks and guidance; our aim is to raise awareness of these and guide repository managers to a subset which is relevant in managing the preservation of crystallography data.

2. Crystallography Data

Crystallography is the sub-discipline of chemistry concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal through the analysis of diffraction patterns obtained from X-ray scattering experiments. Although there are several types of crystallography (chemical, protein, powder etc.) this normally involves several stages which, in broad terms, can be characterised as: data collection; data processing; data workup and publication. Typically, in terms of data volumes, raw data is in the order of Gigabytes, derived or processed data is in the order of Megabytes and results data is normally Kilobytes in size. In terms of data formats, it ranges from proprietary (binary) through to highly structured dictionary defined text.

Over the years there has been a phenomenal growth in the amount of data generated from crystallography experiments; 40 years ago a PhD student would determine 2-3 structures for a thesis - this can now be easily achieved in a single day. However, only a small proportion of the data generated is widely and easily accessible; it is estimated that less than 20% of the crystal structures determined are eventually published [10].

In terms of current practice, the crystallography community takes a relatively organized approach to the management of their derived and results data since crystallography data tends to be highly structured. The convention is to share and exchange derived, or reduced data whilst access to raw data is normally limited to those directly involved in generating the data. Raw data also tends to be subject to individual working practice. The Crystallography Information File (CIF) is the de facto exchange standard [11]. It is maintained by the International Union of Crystallography (IUCr) which is the learned society representing crystallography; it is a publisher of eight journals and maintains standards for communicating and representing crystal structures. There is an established system for publishing crystallographic data alongside journal articles, largely through publisher mandates; the datasets need to be published at the Cambridge Crystallographic Data Centre (CCDC) - a professional body with an international subject repository for crystal data (Crystal Structure Database or CSD). In addition, the Chemical Database Service (CDS), an organisation funded by the EPSRC, provides federated searching across many chemistry databases. Other major databanks include: an inorganic molecule database in Germany; a metals database in Canada and the Protein Data bank in the US. The Royal Society of Chemistry (RSC) is also a key publisher in the field and Chemistry Central is an emerging Open Access publisher operating a repository to store and link data relating to publications in their journals. Reciprocal Net is a distributed database used by research crystallographers to store information about molecular structures; much of the data is available to the general public. More recent developments such as the CrystalEye [12], developed at the Unilever Centre for Molecular Informatics (University of Cambridge), provide open access to aggregated CIF data through the use of web-crawlers.

We recognise that crystallography data can and currently is, managed at many levels: international (ReciprocalNet; CCDC); national (EPSRC NCS, COD); Regional; Institutional (eCrystals); Departmental (local server); Laboratory (PC) and Researcher (laptop, floppy disks, CDs, DVDs). The aim of the eCrystals Federation project is to enhance the management of crystallography data at the institution level, incorporating data generated in departments, laboratories and by individual researchers or practitioners. The considerable

diversity in laboratory practice needs to be taken into account as well as the heterogeneity in instrumentation and its associated software, much of which uses proprietary file formats.

3. An Exemplar Repository: eCrystals@Soton

eCrystals@Soton [13] is the archive for crystal structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service (NCS). As part of the eBank-UK project, the NCS has built an institutional data repository, to provide open access and rapid dissemination of derived and results data from chemical crystallography experiments, as well as linking research data to publications and scholarly communication [14]. The eCrystals data repository started life as a prototype research data repository with the 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.

This report considers the many aspects of preservation planning with reference to our chosen exemplar repository run by the NCS at the University of Southampton.

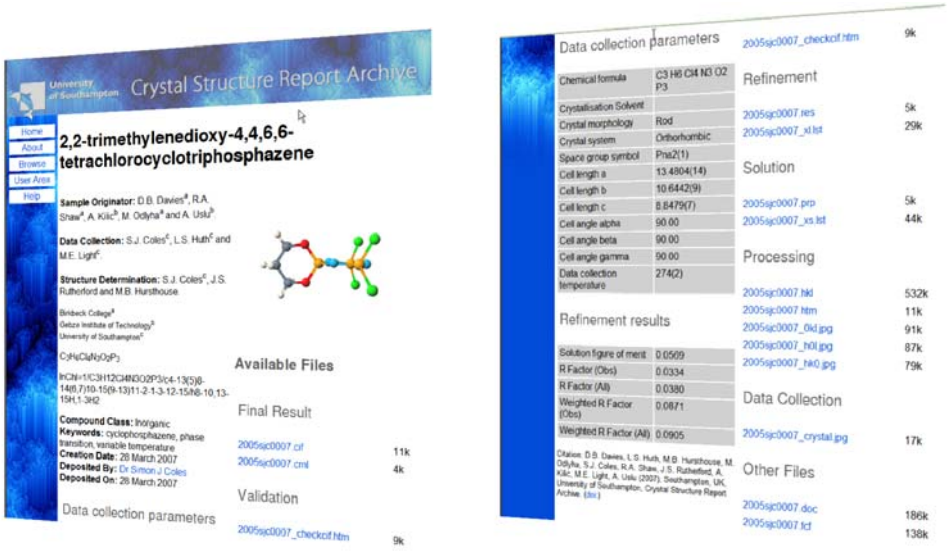

Figure 1: An example Crystal Structure Report in eCrystals [13].

4. Trust and Data Stewardship

Digital preservation is a heavy risk activity due to dynamic and unpredictable developments in economic, social and political terms, as well as in technology over both the short and long term. Digital information can be 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 demand that those entrusted with its stewardship are demonstrably trustworthy to the community that they serve.

> *"Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate"*
>
> [2]Lynch, Clifford A.

[2] *Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age*, ARL, No. 226, February 2003, pp1-7.

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 [15 (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* [16]. 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 confidentiality, intellectual property rights, preservation and the visibility of their deposited content.

The trustworthiness of a content provider depends on several issues, 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.

In January 2007, the Center for Research Libraries (CRL) hosted a 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 [17]:

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, several JISC funded projects have investigated preservation issues within the context of IRs. The SHERPA-DP [18] project examined the practicalities of a disaggregated model. The RepoMMan [19] and REMAP [20] projects examined the embedding of preservation activities within the repository workflow. The PRESERV and PRESERV2 [21] projects have concentrated on incorporating preservation services into the ePrints.org repository software platform [22].

All stakeholders concerned about the longevity of digital data will be interested in whether a repository is capable of identifying and prioritising the threats and vulnerabilities that may impede its activities; whether it is capable of managing those identified risks to mitigate the likelihood of their occurrence and whether the repository is able to establish effective contingencies to alleviate the effects of any risks that do occur. If the repository is able to demonstrate these capabilities, it will engender a status of trustworthiness amongst its user community.

5. Over-arching Tools & Frameworks

A number of tools and techniques are aimed at addressing the criteria for preservation repositories identified in section 4 above. They take differing approaches so that some may be more appropriate than others in certain situations. Note that in planning curation and preservation activities, it is essential to consider these tools within the specific context of a repository; customising and adapting the models; identifying granular functionality; defining roles and responsibilities and building a relevant and appropriate framework of standards and technologies.

5.1 DCC Curation Lifecycle Model

The DCC Curation Lifecycle Model [23], illustrated in Figure 2, provides a graphical, high level overview of the stages required for successful curation and preservation of data from its initial conceptualisation.

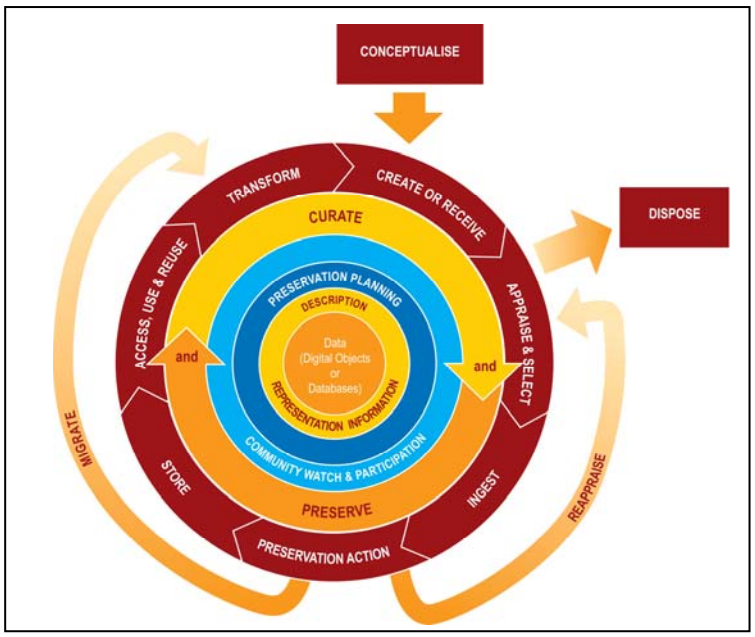

Figure 2: The DCC Curation Lifecycle Model [23]

The model can be used to plan activities to ensure all necessary stages are undertaken, each in the correct sequence. It enables the mapping of granular functionality; definition of roles and

responsibilities; building frameworks of standards and technologies; identification of any additional steps required as well as ensuring adequate documentation of processes and policies. Activities undertaken at each lifecycle stage influence the ability to manage and preserve materials in subsequent stages. The model splits the processes into those that are:

- Full lifecycle stages (*Description and Representation Information; Preservation Planning; Community Watch and Participation; Curate and Preserve*)
- Sequential actions (*Conceptualise; Create or Receive; Appraise and Select; Ingest; Preservation Action; Store; Access, Use and Reuse; Transform)*
- Occasional actions (*Dispose; Reappraise; Migrate)*

5.2 OAIS and other Standards

Within the preservation community, the Reference Model for an Open Archival Information System (OAIS) (ISO 14721:2003 [24]), has established itself as an important standard, influencing: the development of preservation metadata (PREMIS [25]); architectures and systems design of repositories [26]; and conformance criteria for archival repositories [27]. It identifies the environment within which an OAIS operates as well as its basic functions within the context of producers, consumers and the management of data. The Model can be used as a planning tool to facilitate the design of a system which is both sustainable and viable. It can help to ensure that all issues and responsibilities (both internal and external to the repository) are addressed at the planning stage; also that policies, guidelines and agreements exist which will embed preservation into an organisation's workflow; as well as aiding in the identification of roles and responsibilities in the operation of a repository. One of the characteristics of the Reference Model is that it highlights the fact that curation and preservation are continuous processes which may require attention into the indefinite future; hence the model's emphasis on continual monitoring of the environment within which the repository operates.

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 from Producers.
- Obtaining sufficient control of the information to ensure its long-term preservation.
- Determining the "Designated Community" (DC). This is a set of stakeholders and users served by the OAIS. The designated community may be composed of multiple user communities and is subject to change over time.
- Ensuring that information is understandable by the DC without the assistance of the experts who originally produced the information.
- Following documented policies and procedures to ensure that the information is preserved against all reasonable contingencies to enable the information to be disseminated as authenticated copies of the original or traceable as the original.
- Making the preserved information available (dissemination).

Following the OAIS Functional Model will ensure that all the major components required for a successful repository architecture are included: Data Objects are appropriately ingested, stored and managed; administrative procedures are in place for the overall operation of the archive; planning for preservation takes place; including migration planning, software decisions, implementing standards and creating ingest methodologies; so that Data Objects continue to be accessible to the DC over considerable periods of time.

Following the OAIS Information Model will ensure that the necessary supporting information (metadata), to enable effective control and preservation of a Data Object, is collected or created and that any information needed to interpret a Data Object (Representation Information) is also collected and assigned appropriately. Representation Information (RI) is a very broad concept, encompassing any information required to process, render, interpret,

understand and use data. For example, it may be a technical specification, or a data dictionary or a software tool. Issues relating to recording and maintaining RI for crystallographic data are further elaborated in a sister report [28].

Whilst the OAIS Reference Model deals well with data that has been ingested into a repository, it is somewhat lacking on pre-ingest activities. To fill this gap, it is worth investigating the ISO 20652:2006 - Producer-Archive Interface Methodology Abstract Standard or PAIMAS [1] which covers activities such as initial contact, feasibility studies, scope, Submission Information Package (SIP) definition, submission agreement, transfer conditions, access restrictions, delivery, SIP validation and follow-up action with the producer of the information.

Standards to ensure the quality 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; and the information security standard (ISO 17799) which provides a framework for implementing security requirements and quality management (ISO 9001).

5.3 Audit and Certification Instruments

Longevity 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.

Whilst the audit process can be used 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 planning and developmental stages of a repository, facilitating organisational self-awareness and engendering trust from depositors, users, funders and other stakeholders. 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.

5.3.1 TRAC Checklist

The 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* [16]. The report defined the characteristics of a trusted digital repository; listed the 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. This 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. The checklist was made available in the form of a draft for public comment in 2005.

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 and several pilot audits, this work culminated in the latest checklist, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC) published in March 2007 [27]. The checklist has been designed to help institutions objectively evaluate responsibilities against capabilities and identify potential risks to digital content held in repositories and other archives.

A major difference between the RLG-NARA checklist and TRAC is the requirement for documentary evidence relating to various issues, in particular policy and sustainability. 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*. Within each of these categories are various subsections and under the subsections are the 84 criteria themselves.

The TRAC appears 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. However, even a cursory glance at the checklist and its associated criteria is likely to raise awareness of the many threats, vulnerabilities and risks to the long-term survival of digital data.

5.3.2 DRAMBORA Risk Assessment

The *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) toolkit [30] for repository self-audit takes a risk-analysis approach which is complementary to the checklist approach adopted by other audit and certification instruments. A prominent difference between TRAC and DRAMBORA is that the former is well placed for undertaking external audits (with a view to certification), while the latter concentrates more on self-assessment.

The toolkit results from the experience gained through undertaking pilot audits by the DCC based on the RLG-NARA checklist and the NESTOR catalogue [31]. It places emphasis on documented evidence and the assessment and management of risks as critical factors in determining the trustworthiness of a repository. Digital curation is seen to be about taking organisational, procedural, technological and any other uncertainties and transforming them into manageable risks – can the repository be trusted to deliver an 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. The process has six stages; some with multiple sub-tasks (see Table 1).

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. As a result of 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 its intrinsic activities and assets
- 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 (optional)

Stage	Sub-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 related to activities and assests	- Identify risks associated with activities and assets of your repository
5. Assess risks	- Assess the identified risks
6. Manage risks	- Manage the risks identified

Table 1**:** DRAMBORA Risk Assessment Stages [30]

5.3.3 PLATTER Toolkit

The *Planning Tool for Trusted Electronic Repositories* (PLATTER) [32] provides a basis for a digital repository to plan its goals, objectives and performance targets in a manner which will contribute to the repository establishing trusted status amongst its stakeholders. PLATTER is not in itself an audit or certification tool but is rather designed to complement existing audit and certification tools by providing a framework which will allow new repositories to incorporate the goal of achieving trust into their planning from an early stage. The tool focuses only on the process by which the repository organization sets and manages its objectives. The management of the process of implementing these objectives, encompassing such widely disparate areas as finance, human resource management, software and hardware planning, data warehousing etc. is too large an area to be covered by any single document and will typically require input from a range of subject experts.

As a first step in the planning process, PLATTER requires a repository to answer a questionnaire which characterises the repository relative to other repositories and which can be used to determine how and whether the goals and objectives identified are to be realised in a given organisation. The PLATTER process is centred on a group of *Strategic Objective Plans* (SOPs) through which a repository specifies its current objectives, targets, or key performance indicators in those areas which have been identified as central to the process of establishing trust. The intention is that the SOPs should be living documents which evolve with the repository.

5.3.4 Data Seal of Approval

The 17 guidelines for the Data Seal of Approval [33] issued by the Data Archiving and Networked Services (DANS), an institute of the Royal Netherlands Academy of Arts and Sciences, are a distillation of several audit and certification instruments such as those examined above and can be regarded as being on a par with the data stewardship guidelines offered by the Research Information Network (RIN) [3]. Although they refer specifically to the application and verification of quality aspects with regard to creation, storage and reuse of digital research data in the social sciences and humanities, the criteria are equally applicable to other research data.

The objective of the data seal of approval is to safeguard high-quality and reliable processing of research data for the future. The seal of approval gives researchers the assurance that their research results will be stored in a reliable manner and can be reused; it provides research sponsors with the guarantee that research results will remain available for reuse; enabling researchers to assess research data to be reused; and allows data repositories to archive and distribute research data efficiently. Digital research data are required to meet five quality criteria:

1. The research data can be found on the internet.
2. The research data are accessible, while taking into account ruling legislation with regard to personal information and intellectual property of the data.
3. The research data are available in a usable data format.
4. The research data are reliable.
5. The research data can be referred to.

The associated guidelines relate to the implementation of these criteria and focus on three operators: the *data producer*, the *data repository* and the *data consumer*:

- The data producer is responsible for the quality of the digital research data.
- The data repository is responsible for the quality of storage, availability and management of the data.
- The data consumer is responsible for the quality of use of the digital research data.

6. Components of Preservation Planning

One of the obstacles to proactively undertaking preservation and curation activities identified by the interim report of the Blue Ribbon Task force on Sustainable Digital Preservation and Access is a fear that digital access and preservation is too big a problem to take on [34]. In this section we attempt to breakdown the various components of preservation planning to make the process more manageable.

It is increasingly apparent that repository managers are required to have multiple skills, including those in Information Technology (IT) and digital preservation issues. In addition, for those tasked with managing research data, it is crucial to have domain knowledge or expertise in understanding the discipline, the data and its associated community of users. Of course, the level of expertise in each area will depend on the repository manager's own personal experience. Figure 3 illustrates the major components of a generic digital repository and the skills required of repository managers. The aim of this section of the report is to improve the awareness of repository managers with regard to the breadth and multi-faceted nature of preservation issues.

It is beneficial to create an explicit strategy for the stewardship and preservation of digital data. A distillation of the tools, frameworks and guidance in section 5 above, together with a consideration of the ten desirable characteristics of long-term repositories in section 4 suggest that there are a number of common aspects to preservation planning for repositories. Below

we consider these from a pragmatic point of view for the institutional repository manager. The Digital Preservation Coalition's handbook, *Preservation Management of Digital Materials* [35] provides a very useful and practical approach to addressing many of the components identified below.

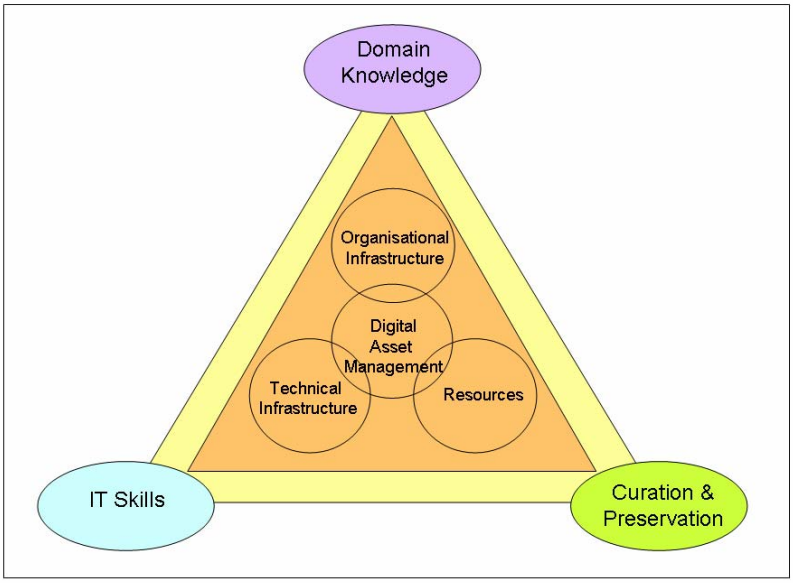

Figure 3: Repository manager skills and components of a digital repository

The process of preservation planning should be driven by an organisation's needs and priorities in relation to its research data, collections and users. One of the key tasks therefore, is to establish the current situation, in terms of what resources the content is comprised of; the media upon which it is stored; the condition of those media; storage facilities and metadata for access and retrieval. A tool such as the Data Audit Framework (DAF) [36] is useful for this purpose. The DAF facilitates the identification and classification of assets; an assessment of how data is currently being managed and recommendations for improvements.

It is pertinent to initially define a high-level strategy identifying organisational aims and objectives with respect to the management of research data and its use, reuse, retention, curation and preservation. This should be followed by a consideration of the operational aspects comprising procedures and activities on a day-to-day basis.

6.1 Analyse Data & Associated Workflows

Critical to developing an effective preservation plan is a thorough understanding of the data as well as the workflows and processes involved in generating it. It is clear that processes and workflows in each crystallography laboratory differ considerably [37,38]. A key requirement is an understanding of the file formats in use as well as the inter-relationships between processing software and data files.

Case Study: eCrystals@Soton

An analysis of the work processes at the NCS (see below) indicates that crystal structure determination involves a near complete digital workflow. It highlights the various file formats in use and the relationships that exist between them. It is also apparent that specialised data formats are tightly linked to specific analysis and processing tasks. Broadly, there are three categories of data involved:

Raw data – images (JPEG) and proprietary formats (.kcd)

Derived data – processed data in the form of de facto community standard formats (.hkl, .prp, .res, .lst etc.)
Results data - crystal structures in standard formats (.cif, .cml, .mol)

Procedures at the NCS indicate that a number of well-defined, sequential stages are readily identifiable and result in a workflow as shown in Figure 4. At each stage, an instrument or computational process produces an output, saved as one or more data files which provide input to the next stage. The output files vary in format, they range from images to highly-structured data expressed in textual form; the corresponding file extension names are well-established in the field. Some files also contain metadata, such as validation parameters, about the molecules or experimental procedures.

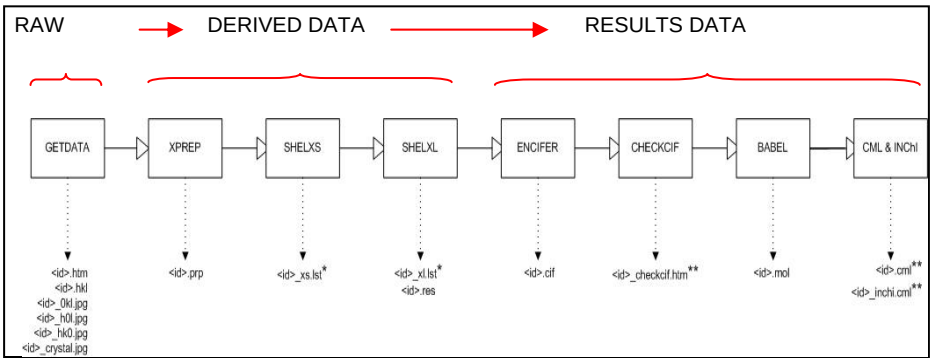

Figure 4: Workflow model of the EPSRC UK National Crystallographic Centre (NCS)

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 in Figure 2 above, 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 [11] is used within the community as an interchange format and is supported by the IUCr – the 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 also capable of describing the whole experiment and modelling processes. 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 both syntactically and for crystallographic integrity; it is made available as an open web service by the IUCr.

Another type of file format included in the Final Result is a Chemical Markup Language (CML) encoding [39]. The CML file is translated from the CIF and introduces complimentary semantic information such that between them they provide a complete description of the molecule as well as its chemistry. The {.mol} file is a useful intermediate format for producing the InChI [40], a unique text identifier that describes molecules, and is generated from the {.cif} file. These file format conversions are performed according to well defined standards using the OpenBabel [41] software obtainable from SourceForge.

The data collection, processing and solution stages involve the major 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 used in the experiment. 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 [42] 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. As shown in Figure 1, a 3D graphical rendition of the molecule (jmol) is also generated; this can be rotated interactively on the eCrystals website.

6.2 Evaluate Preservation Requirements

The next stage is to evaluate the long-term value of the data (appraisal) and the consequent preservation requirements of the data in relation to the repository's envisaged user community. This involves an examination of the lifecycle of the data from its inception to the end of its life, how it is used and over what periods of time in order to establish timeframes and long-term data management requirements. The DCC Curation Lifecycle Model [43] serves as a useful reference.

Case Study: eCrystals@Soton

At the NCS, all data generated through crystallography experiments is considered to have long-term value, however there is a difference in the long term value of derived and results data. Derived data has more value initially, but this reduces with time as the structure becomes established as a trusted study. The results data can be reused indefinitely into the future. Raw data, on the other hand, is interesting in that it is well known that in some cases contemporary extraction software is not capable of capturing everything, so that it is worth bearing in mind that in the future new algorithms may come along and enable further insight into the crystal structure.

For journal publications that report and link to crystal structure determinations presented in the repository, it is important to satisfy both publishers and the public that eCrystals@Soton will have the same stability and longevity as journal publications.

In addition, since the NCS is a national service there is felt to be an obligation to retain data, although exactly how long for is not currently specified. NCS offers two type of experimental service to its users:

1. Full determination, where NCS generates raw data and works up derived data into results. This is deposited in eCrystals (generally initially embargoed for 3 years).
2. Data Collection only, where NCS collects the raw data and turns it into the first stage of derived data. This derived data is then sent to users and they work it up into results. None of the derived or results data is deposited in eCrystals.

The future plan is to use eCrystals for case 2 above as well as case 1, so that the NCS will deposit first stage derived data; allow the user to pick this up, turn it into result data and deposit the final result into eCrystals.

6.3 Define a Preservation Policy

Sound policy development combined with effective working practices and procedures is essential to effective digital preservation programmes; however it is important to remember that the repository's preservation policy also needs to take account of the priorities of its larger organisation or institution with respect to crystallography data, as well as to make sure that the repository's policies are viable in the environment in which it is operating.

Policies are vital for ensuring compliance with procedural and legal requirements within an organisation. By clearly defining a set of procedures, roles and responsibilities, policies help to promote transparency and accountability. Additionally, policies provide an overall cohesion within an organisation and offer guidance for best practice.

Plans for managing and maintaining research data will vary considerably depending on the organisational context within which they are being formulated. A recent JISC funded study, *Digital Preservation Policies Study* [44], aims to provide an outline model for digital preservation policies, whilst the OpenDOAR Policies Tool [45] allows repository managers to quickly generate policy statements relating to

- Metadata (access and reuse)
- Data (access to and reuse of full items)
- Content (repository type; type of material held; principal languages)
- Submission (eligible depositors; deposition rules; moderation; content quality control; publishers' and funders' embargos; copyright policy)
- Preservation (retention period; functional preservation; file preservation; withdrawal policy; withdrawn items; version control; closure policy)

Additional advice on policy issues is also available in the form of a briefing paper by the Repositories Support Project [46] and a guidance document published by the JISC-funded DISC-UK DataShare Project (2007-2009). The *Policy-making for Research Data in Repositories: A Guide* [47], is intended to be used as a decision-making and planning tool, as more institutions expand their digital repository services into the realm of research data to meet the demands of researchers who are themselves facing increasing requirements from funders to make their data available for continuing access.

Case Study: eCrystals@Soton

The eCrystals repository at present has basic information with regard to the contents of the repository, use of the data and citation thereof [48,49]. A formal preservation policy is not currently in force and is an area that needs to be addressed.

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.

6.4 Formulate a Preservation Strategy

Technological obsolescence of hardware, software and file formats is one of the greatest threats to digital information. Due to the dynamic and unpredictable nature of technology, digital information is extremely vulnerable, being susceptible to undocumented change, modification and technological obsolescence, so that data can become inaccessible within a very short period of time. In addition, digital data is characterized by the fact that it requires rendering or visualization applications in order to make is accessible to humans as well as details of the semantics associated with the data to make it understandable and usable. At the heart of any curation and preservation strategy is the need to maintain the security, integrity and authenticity of the data. Reliable reuse of digital data is only possible if data is curated in such a way that these aspects are attended to.

Ensuring that digital data remain accessible and reusable over time requires the implementation of proactive, scalable and sustainable preservation strategies. To be of

greatest effect, preservation issues must be considered from the point of creation of the data and throughout its entire life-cycle. It should be noted that coping with technological obsolescence is a huge task in itself.

A variety of techniques have been proposed and explored over the years to combat the effects of rapidly changing technologies and media degradation. Broadly speaking, the majority of these strategies are aimed at preserving access to the byte-stream of digitally encoded data. Here, we would like to make a distinction between digital data and the information or intellectual content that it carries. Whilst merely ensuring access to the bits and bytes of a digital asset may not be of much use in terms of understanding the content and its reusability, it is nevertheless a necessary precondition of such activities.

It should be recognized that no single strategy of those suggested is likely to be appropriate for all types and varieties of digital research data, circumstances and organizations. In practice, a strategy that combines elements from several of the possible approaches is likely to provide the most effective solution. The techniques that are aimed largely at bit-preservation include: bit-stream copying (in essence, backups), refreshing, the use of durable media, digital archaeology and replication [50]. Such techniques can be considered on a parallel to the provision of a secure technical environment coupled with sound data management practice such as regularly re-assessing and transitioning digital data to new storage media and maintaining multiple copies of the data off-site as well as on differing computer platforms. To ensure that data is secure and unaltered, it is recommended that it is stored in an effective Information Security Management System (ISMS). A DCC Standards Watch Paper [51] provides coverage of the ISO 27000 (ISO 27K) Series on Information Security Management which specifies how to implement a successful ISMS.

However, curation and preservation of digital data for contemporary and future use involves several additional aspects over and above those for bit-preservation, including ensuring the quality and authenticity of the data and undertaking format migration for long-term access. Strategies aimed at preserving access to the information content and providing functional preservation include: technology preservation, analogue backups, migration, normalisation (with a reliance on standards), emulation and encapsulation [50].

Managing format migration is of paramount importance; selection of a format for preserving access to research data will be dependent on what aspects of the resource will be required in the future or its *significant properties* [52]. In addition, it should be noted that format migration may involve changes to the data such that Intellectual Properties Rights need to be taken into account. Several registries (GDFR [53], PRONOM [54]) and tools (JHOVE [55], DROID [56]) are available for identifying and validating file formats –although at present they do not contain details of current crystallography file formats.

A particular strategy concerned with mitigating the effects of technology evolution is the use of the OAIS concept of Representation Information (RI). This is essentially any information that is required to render, process, visualize, extract meaning from and use data. In addition, an OAIS Archival Information Package (AIP) comprises both RI and Preservation Description Information (PDI) and is in effect a form of encapsulation collecting together all the information relevant to the preservation, interpretation and reuse of digital data. An investigation into RI for crystallography data is presented in a sister report [28].

It is important to remember that not all preservation services need to be catered for internally and that it may be pertinent to evaluate options for out-sourcing certain aspects (e.g. using Amazon S3 for storage [57]).

Case Study: eCrystals@Soton

eCrystals is constructed on the ePrints.org repository software platform [22] (version eprints-3.0.3-rc-1) which has been customised specifically to cater for crystallography data. The repository makes use of the institutional computer infrastructure and network.

The current strategy is such that raw data is archived and preserved in perpetuity off-site, at the Atlas Data Store (based at the Rutherford Appleton Laboratory) since it is not regularly accessed; in the meantime derived and results data are made available through the eCrystals repository for validation and re-analysis over the Web. Raw data on the Atlas Data Store goes back to 2002, whilst raw data from 1998-2002 is stored on USB disks stored in the NCS laboratory (migrated from CD's written at the time of generation).

The eCrystals server is managed by a part-time systems administrator with primary training in crystallography. Backups of the repository are kept within the Chemistry department, in another building to where the main server is housed. At the present time the repository comprises 4 terabytes of data; the associated metadata can be exported using a METS profile [58] to allow ingest to an alternative repository platform. The use of OAI-ORE [59] for packaging crystallography data for interoperability purposes is currently under investigation.

In addition, experiments are underway to provide an institutional solution to the storage of raw data in the form of Sun's Honeycomb platform, but are likely to be discontinued in the face of Sun's withdrawal of support for the hardware beyond 2013.

Raw data is proprietary in nature since it is dependent on specific instrumentation; however beyond initial processing, the data ends up in a normalised, de facto community standard format which is portable and usable by other crystallographers (see section 6.1). At the end of a crystal structure determination, the results data is in the form of a standard CIF file. A considerable amount of quality and validation checking is performed prior to data files being ingested into eCrystals.

In the case of crystallography data, it is clear that processing software plays a very important part in crystal structure determination (see section 6.1). In particular, software such as the SHELXL/S suite of programs, as well as those for checking and validating CIF files (checkCIF) may also need to be curated and preserved.

Several preservation strategies beyond bit-preservation have been considered including emulation and migration. However, continuous migration of formats carries the risk of data corruption as well as requiring considerable effort and resources given the amount of data in eCrystals (4 terabytes); although this could be mitigated to some extent, by a strategy of *migration on demand*. Emulation has also been considered as a possibility, but again would require resources currently beyond the means of the repository.

6.5 Record Preservation Metadata

Fundamental to preserving and curating digital information, is the recording of adequate and appropriate documentation or metadata. The OAIS Reference Model 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); 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.

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 different 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. A sister report describes the formulation of preservation metadata for crystallography data [60].

Whilst the exact metadata to be recorded is dependent on the specific preservation strategy in force, there is some consensus on a core set of preservation metadata (PREservation Metadata: Implementation Strategies (PREMIS) Data Dictionary [25]):

> *"things that most working preservation repositories are likely to need to know in order to support digital preservation"*

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: *Implementing the PREMIS data dictionary: a survey of approaches*, A Report for the PREMIS Maintenance Activity [61]; *Preservation Metadata for Institutional Repositories: applying PREMIS* [62] and *Review of metadata standards in use by SHERPA DP repositories* [63].

According to Caplan [64], 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 [54] or the GDFR [53]. For automated extraction of format information, tools such as JHOVE [55] and DROID [56] 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.

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.

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).

Case Study: eCrystals@Soton

At present eCrystals uses the eBank-UK Metadata Application Profile [65] from which much of the necessary curation and preservation metadata is absent. Broadly speaking, the profile records the following:

Simple Dublin Core
- *Crystal structure*
- *Title (Systematic IUPAC Name)*
- *Authors*
- *Affiliation*
- *Creation Date*

Qualified Dublin Core (for additional chemical metadata)
- *Empirical formula*
- *International Chemical Identifier (InChI)*
- *Compound Class and Keywords*

Each crystal structure report in the repository is also assigned a Digital Object Identifier so that the entry may be referenced in any future publication. A rights and citation statement is also available on the eCrystals website [49].

In developing the eCrystals repository, NCS discovered that most crystallographers are wary about making their data immediately available for open access; this resulted in the formulation of an embargo scheme whereby data that is up-loaded is initially stored in a closed part of the repository and re-assessed for open access 3 years later. This private part of eCrystals is currently used as a comprehensive laboratory management and data archival system, to which all completed and validated crystal structure determination outputs are uploaded.

A cursory evaluation of the PREMIS Data Dictionary suggested a set of top-level Semantic Units which are likely to be of importance for the eCrystals data repository [66]. However, work is currently in progress with regard to the recording of preservation metadata in eCrystals and the associated federation [60].

6.6 Model Costs

Identifying and securing the requisite resources to maintain a digital repository is a critical task, but this should be done on a cost-benefit analysis basis, to quantify the value of preserving research data as a means of securing funding for digital preservation activity. In addition, it is of paramount importance that resources are adequate to meet the aims of any preservation policy and strategy that may be in force; consequently it is necessary to undertake comprehensive modelling to ensure that all costs are factored in. Costs will necessarily vary depending on:

- The repository's policy with regard to the long-term management of research data
- The nature of the data and decisions with regard to what and how much to preserve
- The significant properties or particular characteristics of the data that need to be retained
- The specific choices made with respect to the details of the repository's preservation strategy

However, cost modelling is not an exact science and there will, in all probability, be unforeseen expenses such as a costly migration of a rare or complex file format, or perhaps IPR issues related to data migration.

In broad terms, costs can be categorised into three types: start-up, on-going and contingency. The costs of setting up a repository should include all capitol costs associated with the required technical and procedural infrastructure. One technique for ensuring a comprehensive coverage of direct operating costs is to categorise them on the basis of the functions identified in the OAIS Reference Model: ingest, data management, storage, preservation and dissemination. Staff costs tend to represent tend to represent the greatest proportion of total costs and will only decrease with the automation of various tasks. However, there may be limits to automation; for example it is well known that recording quality and checked metadata is an expensive and labour intensive business. On the other hand, per submission and per item costs may well decrease as the amount of content in a repository grows.

Indirect operating costs or overheads, which include general administrative and support services, also need to be taken into account; and it is advisable to maintain a certain amount of contingency funding for unexpected occurrences or for example out-sourcing certain functions if they are no longer viable internally to an organisation.

In predicting or forecasting costs, it is as well to take inflation into consideration, since present day costs will not be the same as those in the future. Preservation services should be

regarded as part of the infrastructure required for effective data management, and as such should be included in any data management plans and costs thereof.

The costs incurred in undertaking curation and preservation activities have always been recognised as a major factor in their uptake and several investigative models and studies have been commissioned [67]. The LIFE and LIFE2 projects [68] have developed a methodology to model the lifecycle of digital data including: creation, selection, ingest, storage, retrieval and preservation and the costs associated with each stage.

A recent study undertaken by Beagrie et al. [69] focuses on developing a framework and guidance for determining medium to long-term costs to Higher Educational Institutions of the preservation of research data. The study reviews the LIFE model and applies the cost framework developed to eCrystals@Soton as one of several case studies (phase two of this study has recently started and eCrystals@Soton is once again a case study).

Finally, the costs of regularly assessing and evaluating preservation plans should not be forgotten and should include the monitoring of standards, technology, file formats and developments within the user community as well as staff training.

Case Study: eCrystals@Soton

At present the running and maintenance of the repository is dependent largely on gaining regular funding by the NCS which is financed by the EPSRC on a three yearly rolling grant basis. The budget currently explicitly covers the cost of raw data storage at the Atlas Data Store, but not that of derived and results data – an issue which is currently under review.

A thorough and detailed cost analysis of eCrystals@Soton is provided in the study by Beagrie et al. [69], resulting in an average cost for preserving a results dataset as £2.15 and the average cost for out-sourcing the archiving of a raw dataset as £1.48. For details with regard to factors which have been included and those omitted, the reader is referred to the study itself.

6.7 Plan for Sustainability

Sustainability planning is an important factor in gaining the trust of data depositors and users. Sustainability of crystallography data can be considered at several different levels including at a business model level; an individual repository level; an institutional level and at a crystallography community level. Here, we focus on the repository and community levels.

Sustainability plans at a repository level should include forward planning for transitions in data stewardship. In the event of the repository becoming unviable what will happen to the content data? If the data is to be transferred to another custodial environment, ensure that it meets the requirements of its new home and arrange the necessary agreements from the new steward.

The eCrystals Federation project can be viewed as being a part of community supported sustainability planning. Within the federation, models such as those of LOCKSS (Lots of Copies Keeps Stuff Safe) and/or CLOCKSS (Controlled LOCKSS) [70] could be considered as a means of supporting individual repositories to provide off-site back-up and recovery services. Additional activities at a community level could include engagement with and gaining support from the IUCr and the Royal Society of Chemistry as well as seeking maintenance and open access of critical file formats and software such as: the Crystallography Information File (CIF); work-up software e.g. XPREP; SHELXS/L; ENCIFER; checkCIF, BABEL; and advocacy for the export of raw data from instrumentation in the form of the IMG CIF file format [11]. Collaborative measures could also be taken with respect to the capture of RI for crystallography data [28]; consensus on a crystallography metadata

Application Profile [60] and the automation of metadata generation, extraction and maintenance.

Case Study: eCrystals@Soton

At a repository level, although the content of eCrystals is regularly backed-up there appears to be no formal plan for the management of the data in the event that the repository becomes unviable (although negotiations are in progress with the IUCr with regard to this issue). However, the metadata is capable of being exported to another repository platform using a METS profile [58] and the use of OAI-ORE [59] for packaging crystallography data for interoperability purposes is currently under investigation.

At the community level, the Director of the NCS is proactively involved in many initiatives related to the open access, sharing and long-term sustainability of crystallography data.

6.8 Regular Evaluation and/or Self-Assessment

For a long-term repository it is beneficial to have regular evaluations or self-audits, which verify periodically the proper functioning of records, management procedures and systems as well as the authenticity and reliability of the research data. Such monitoring is also useful in building up a profile of the repository over time in the face of a continuously changing environment. A self-assessment could be undertaken 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.

Such evaluations also play a major part in gaining the trust of the repository's user community; it is necessary to demonstrate compliance with the long-term repository criteria outlined in section 4. This can be achieved in a formal manner through the use of a tool such as DRAMBORA or TRAC. Compliance with the relevant criteria needs to be demonstrated through documentation (evidence), transparency (open examination of the evidence), adequacy (degree to which the evidence meets the vision and goals) and measurability.

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. DRAMBORA takes a more quantified approach to assessing repositories and would therefore work best for an established repository looking for self-assessment. TRAC, on the other hand is more open-ended and exploratory, taking into account vision, goals and plans and therefore more suited to repositories with an established long-term archival and preservation mandate, such as a national archive.

As well as under-going practical tests to verify the recoverability of data, it is also important to regular revisit and revise preservation policies, strategies and plans to keep them up-to-date with the repository's mission.

Case Study: eCrystals@Soton

The eCrystals repository has been in existence since 2005; in 2008 it became NCS policy to store in the repository all derived and results data generated by the national service. However, there are no formal records of evaluation, testing or self-assessment. Given that the repository is currently maintained by a systems administrator working at 0.5FTE, the opportunity for a thorough evaluation has so far been limited.

7. Conclusions

Guaranteeing the long-term safety and accessibility of fragile digital research data involves a substantial commitment – a commitment which must remain on-going and continuous, and

one which we cannot afford to suffer gaps in attention. Our exemplar repository has demonstrated that even for a repository maintained by a national service there is a long way to go to achieve adequate curation and preservation of crystallography data.

This report has attempted to illuminate the many aspects and factors in taking on custodial responsibility for research data, and crystallography data in particular, in the hope that repository managers will be able to make informed decisions with regard to how best to manage the valuable research data that they are entrusted with, given the resources at their disposal.

Repository managers are now expected to be familiar with curation and preservation issues for the content that they manage –a task which often seems rather daunting. We have tried to break-down the preservation planning process into more manageable components with a practical focus. The recently announced KeepIt! Project [71] will be undertaking several case studies with the specific aim of developing practical preservation solutions for each repository.

Factors that influence preservation planning are very wide-ranging and need to be assessed in terms of their relevance to a particular context and situation. For example, changes in user expectation or government policy may need to be taken into account, or variations in finances may affect what can be achieved or new technology may lead to changes in procedures. For this reason, we have identified several useful frameworks and tools which can be used in a methodical manner to aid in the development of a comprehensive preservation plan. It should be borne in mind that it is important to have a flexible plan that can be revisited and revised regularly to allow change to be managed.

References

1. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, Report of the National Science Board (Draft), May 2005, http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf
2. 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
3. *Stewardship of Digital Research Data – A Framework of Principles and Guidelines*, Research Information Network, January 2008, http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf
4. *The UK Research Data Service Feasibility Study*, UKRDS Report and Recommendations to HEFCE, December 2008, http://www.ukrds.ac.uk/
5. Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, JISC Consultancy Report, June 2007, http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf
6. 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/
7. European Initiative to Facilitate Access to Research Data, March 2009, http://www.dini.de/service/nachrichten/nachricht/x//european_initiative_to_fa/
8. *Digial Curation Centre Statement and Charter of Principles*, December 2009, http://www.dcc.ac.uk/charter/
9. eCrystals Federation Project, http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page

10. Allen, F. H., *High-throughput crystallography: the challenge of publishing, storing and using the results*. Crystallography Reviews, 10, pp3-15 (2004).
11. CIF -The Crystallographic Information File, http://www.iucr.org/iucr-top/cif/
12. CrystalEye, Unilever Centre for Molecular Informatics, University of Cambridge, http://wwmm.ch.cam.ac.uk/crystaleye/
13. The Crystal Structure Report Archive –eCrystals Data Repository, http://ecrystals.chem.soton.ac.uk
14. Monica Duke, Michael Day, Rachel Heery, Leslie A. Carr, Simon J. Coles *Enhancing access to research data: the challenge of crystallography* JCDL 2005 Digital Libraries: Cyberinfrastructure for Research and Education, Denver, Colorado, USA June 7-11, 2005
15. *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
16. *Trusted Digital Repositories: Attributes and Responsibilities*, An RLG OCLC Report, May 2002, http://www.rlg.org/legacy/longterm/repositories.pdf
17. *Ten Core Requirements for Digital Archives*, Centre for Research Libraries, January 2007, http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92
18. JISC SHERPA-DP Project, http://www.sherpa.ac.uk/projects/sherpadp.html
19. JISC RepoMMAN Projecet, http://www.hull.ac.uk/esig/repomman/
20. JISc REMAP Project, http://www.jisc.ac.uk/whatwedo/programmes/preservation/remap.aspx
21. JISC PRESERV projects, http://preserv.eprints.org/
22. EPrints Digital Repository Software, http://www.eprints.org/software/
23. DCC Curation Lifecycle Model, 2008. http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf
24. 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%20model%22
25. PREMIS Data Dictionary for Preservation Metadata version 2.0, Preservation Metadata Maintenance Activity, 2008, http://www.loc.gov/standards/premis/
26. Giaretta, D., T*he CASPAR Approach to Digital Preservation*, International Journal of Digital Curation, Vol. 2 (1) (2007)
27. *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?l1=13&l2=58&l3=162&l4=91
28. Manjula Patel, Representation Information for Crystallography Data, WP4, eCrystals Federation Project, http://wiki.ecrystals.chem.soton.ac.uk/images/e/e1/ECrystals-WP4-RI-090519.pdf
29. Consultative Committee for Space Data Systems ,*Producer Archive Interface Methodology (PAIMAS)*, ISO 20652:2006, http://public.ccsds.org/publications/archive/651x0b1.pdf
30. *DRAMBORA -Digital Repository Audit Method Based on Risk Assessment*, March 2007, Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), http://www.repositoryaudit.eu/
31. Catalogue of Criteria for Trusted Digital Repositories, Version 1 (Draft), NESTOR Working Group Trusted Repositories - Certification, December 2006, http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_07.pdf
32. *Planning Tool for Trusted Electronic Repositories* (PLATTER), PLANETS Project, D3.2, April 2006
33. *Data Seal of Approval*, Data Archiving and Networked Services (DANS), http://www.dans.knaw.nl/en/data_deponeren/dans_keurmerk/

34. *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*, Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, December 2008, http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf
35. *Preservation Management of Digital Materials – The Handbook*, Digital Preservation Coalition, http://www.dpconline.org/graphics/handbook/
36. Data Audit Framework (DAF), http://www.data-audit.eu/
37. Liz Lyon, Simon Coles, Monica Duke, Traugott Koch *Scaling Up: Towards a Federation of Crystallography Data Repositories*, eBank-UK Project, Phase 3, May 2008, http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/Ebank3report/Ebank3report.pdf
38. *To Share or not to Share, Publication and Quality Assurance of Research Data Outputs*, A Report commissioned by the Research Information Network (RIN), Annex: detailed findings for the eight research areas, June 2008, http://www.rin.ac.uk/node/218/data-publication
39. Chemical Markup Langauage (CML), http://www.ch.ic.ac.uk/rzepa/cml/
40. IUPAC International Chemical Identifier (InChi), http://www.iupac.org/inchi/
41. Open Babel: The Open Source Chemistry Toolbox, http://openbabel.org/wiki/Main_Page
42. SHELX Home Page, http://shelx.uni-ac.gwdg.de/SHELX/
43. Sarah Higgins, *The DCC Curation Lifecycle Model*, International Journal of Digital Curation, Vol 3 (1), 2008, http://www.ijdc.net/index.php/ijdc/article/view/69/69
44. Beagrie *et al*., Digital Preservation Policies Study, JISC, 30 October 2008 http://www.jisc.ac.uk/Home/publications/publications/jiscpolicyfinalreport.aspx
45. OpenDOAR Policies Tool, http://www.opendoar.org/tools/en/policies.php
46. *Repository Policy Framework*, Repositories Support Project http://www.rsp.ac.uk/pubs/briefingpapers-docs/repoadmin-policyv2.pdf
47. *Policy-making for Research Data in Repositories: A Guide*, DISC-UK DataShare Project, May 2009, http://www.disc-uk.org/docs/guide.pdf
48. eCrystals, About, http://ecrystals.chem.soton.ac.uk/information.html
49. eCrystals, Rights & Citation, http://ecrystals.chem.soton.ac.uk/rights.html
50. *Digital Preservation Strategies*, Preserving Access to Digital Information (PADI), National Library of Australia, http://www.nla.gov.au/padi/topics/18.html
51. Sarah Higgins, "*Information Security Management: The ISO 27000 (ISO 27K) Series*" DCC Standards Watch Paper, May 2009, http://dcc.ac.uk/resource/standards-watch/information-security-management/
52. Wilson, A. 2007. InSPECT significant properties report. Technical report, Arts and Humanities Data Service/National Archives
53. The Global Digital Format Registry (GDFR), Digital Library Federation, http://hul.harvard.edu/gdfr/
54. The PRONOM registry, The National Archives, http://www.nationalarchives.gov.uk/pronom/
55. JHOVE -JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/
56. DROID –Digital Record Object Identification, The National Archives, http://droid.sourceforge.net/wiki/index.php/Introduction
57. Amazon Simple Storage Service (Amazon S3), http://aws.amazon.com/s3/
58. Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets/
59. Open Archives Initiative Protocol – Object Reuse and Exchange, http://www.openarchives.org/ore/
60. Manjula Patel, *Preservation Metadata for Crystallography Data*, WP4, eCrystals Federation Project, To appear

61. 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
62. 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
63. Gareth Knight and Kirti Bodhmage, *Review of metadata standards in use by SHERPA DP repositories*, February 2006, http://www.sherpadp.org.uk/documents/wp41-metadata_standards.pdf
64. Priscilla Caplan, *Instalment on "Preservation Metadata"*, DCC Curation Manual, July 2006, http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/
65. eBank-UK: Metadata Schemas, http://www.ukoln.ac.uk/projects/ebank-uk/schemas/
66. Patel,M. and Coles S.: A study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation, eBank-UK Phase 3, Scoping Report, July 2007
67. Workshop on Cost Models for Preserving Digital Assets, DCC/DPC, July 2005, http://www.dpconline.org/graphics/events/050726workshop.html
68. LIFE2 Conference, http://www.life.ac.uk/2/conference.shtml
69. Neil Beagrie, Julia Chruszcz, Brian Lavoie, *Keeping Research data Safe: A cost Model and Guidance for UK Universities,* A report commissioned by the JISC, April 2008, http://www.jisc.ac.uk/publications/documents/keepingresearchdatasafe.aspx
70. LOCKSS –Lots of Copies Keeps Stuff Safe, http://www.lockss.org/lockss/Home
71. JISc KeepIt! Project, http://preservation.eprints.org/keepit/