# Preservation Metadata for Crystallography Data

## JISC eCrystals Federation Project

## WP4: Repositories, Preservation and Sustainability

**Document Details**

| | |
|---|---|
| Author: | Manjula Patel (UKOLN & DCC) |
| Date: | 3rd September 2009 |
| Version: | 1.0 |
| File Name: | eCrystals-WP4-PM-090903.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. WP4 of the project is concerned with the development of approaches to the preservation and curation of crystallography data in open repositories. In terms of the crystallography community, the long-term provision of data is particularly important since structure determination can only be truly repeated or verified when the raw data is available. In addition, the availability of raw data is extremely useful for reanalysis and reprocessing as improved methods for performing these tasks emerge.

Metadata is essentially any information that documents the characteristics and attributes of a resource. The term is often defined as "*structured data about a resource*". A metadata vocabulary supports a wide variety of functions, for example description, identification, discovery, retrieval, rights management and preservation. Metadata is consequently pivotal in the management of all types of resources, helping to ensure that they will survive and continue to be accessible and usable into the future. A structured set of metadata elements is normally organised into a schema, representing a data model and the attributes associated with the entities within it.

We consider that preservation activities should be viewed as an integral part of sound data management practice. As a result, metadata that supports curation and preservation should be embedded into the core metadata for managing crystallography data. However, metadata creation, capture and maintenance is generally regarded as being tedious, time-consuming, subjective and labour-intensive and therefore very costly. Given the critical role that metadata plays in the management and curation of electronic resources it is important to craft the structure and architecture of the metadata correctly from the outset so that adequate and appropriate information is recorded.

We examine work that has already been done in the area of preservation metadata, in particular the influence of the Open Archival Information System (OAIS) Reference Model (PDI – Preservation Description Information) and the PREMIS (PREservation Metadata: Implementation Strategies) working group which has produced a Data Dictionary for Preservation Metadata – a core set of preservation metadata i.e. "*the information most preservation repositories need to know to preserve digital materials over the long-term*".

Preservation metadata is information that supports and documents the digital preservation process. It is sometimes considered a subset of technical or administrative metadata and incorporates:

- Provenance: *Who has had custody or ownership of the digital object?*
- Authenticity: *Is the digital object what it purports to be?*
- Preservation Activity: *What has been done to preserve the digital object?*
- Technical Environment: *What is required to render and use the digital object?*
- Rights Management: *What intellectual property rights must be observed?*

The primary aim of preservation metadata is to support preservation activities; consequently, differing preservation strategies are likely to demand that distinct types of information be recorded. For example, a preservation plan based on migration activities will require different information to that of one based on emulation. Hence, the preservation plans and policies of a particular repository will heavily influence the additional specific metadata that is to be recorded.

The technical aspects of digital curation and preservation are only one facet of a multidimensional problem; curatorial issues further encompass social, cultural, political,

organisational, financial and legal factors as well. Community consensus and the development of standards and guidelines for best practice underpin the longevity, effective management, preservation, sharing and reuse of science data. To this end, a collaborative venture named *Towards an International Data Commons for Crystallography* (TIDCC) emerged as a result of discussions between participants in the TARDIS (The Australian Repositories for Diffraction Images), eCrystals Federation and DataMINX Projects and the Australian Research Council's Molecular & Materials Structure Network (MMSN) in September 2008. The intention of the TIDCC is to develop a community derived metadata schema capable of describing all types of crystallography data related to single crystal diffraction.

We use the notion of an *Application Profile* (AP) in contemplating the metadata required to manage and preserve crystallography data, building on several schemas including that of the eBank-UK AP; the TARDIS schema and the CCLRC Scientific Metadata Model (CSMD). The purpose of an AP is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, whilst retaining interoperability with the original base schemas. This offers the potential for digital materials to be accessed, used and curated effectively both within and beyond the communities in which they were created.

The TIDCC Metadata Application Profile (TMAP), which is presented in an Appendix, was a first attempt at constructing an over-arching AP for crystallography data which would facilitate the exchange of not only metadata, but also the data itself. However, following several meetings is has become apparent that a more effective way forward is to adapt the ICAT data model (a simpler version of the CSMD) and schema to cater for curatorial and preservation activities since ICAT is presently being used by a growing proportion of the science community for managing their data. As of the completion of this report, the work is still in progress; it is expected that the preservation metadata proposed in the TMAP will feed into the new development.

It is clear that the crystallography community recognises the importance of high quality metadata for all the functions that it can support, including the long-term accessibility and reuse of scientific data. Although there is still a considerable way to go along the path to formulating community agreed metadata for the curation and preservation of crystallography data, the work outlined in this report proves that the crystallography community appreciates the benefits and does not lack the motivation to achieve such as goal.

# Acknowledgements

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

The following people were involved in the development of the TIDCC schema: Simon Coles, Richard Stephenson, Manjula Patel, Ashley Buckle, Peter Turner, Antony Beitz, Steve Androulakis and R Brownlee.

# Revision History

| Date | Contributor | Contribution |
|---|---|---|
| 15th October 2008 | M Patel | Setting up and outline of report and associated work |
| 17th February 2009 | M Patel | Re-structuring of contents to take account of Crystallography Data Commons work and proposals for preservation metadata in TIDCC schema |
| 4-14th May 2009 | M Patel | Wrote up section on Preservation Metadata Standards |
| 18th May 2009 | M Patel | Added place holder and notes on Crystallography Data Commons |
| 8-10th June 2009 | M Patel | Added section on eBank-UK AP |
| 2-20th July 2009 | M Patel | Scoping and writing up of sections 1 & 2 |
| 29-31st July 2009 | M Patel | Added section on Application Profiles |
| 15-20th August 2009 | M Patel | Added section on Crystallography Data |
| 25th August 2009 | M Patel | Added section on CSMD |
| 26-28th August 2009 | M Patel | Added section on TIDCC Data Model and Schema |
| 29-31st August 2009 | M Patel | Tidying up of text |
| 1st-2nd Sept 2009 | M Patel | Added sections on Community & Federation issues; Conclusions and References |
| 3rd Sept 2009 | M Patel | Added Executive Summary |

# Contents

# 1. Introduction

Metadata is essentially any information that documents the characteristics and attributes of a resource. The term is often simply defined as "*data about data*", although it is more commonly understood to mean "*structured data about a resource*" that supports a wide variety of operations relating to that resource, for example description, identification, discovery, retrieval, rights management and preservation. Metadata is consequently pivotal in the management of all types of resources, helping to ensure that they will survive and continue to be accessible and usable into the future.

The metadata associated with a resource comprises a set of elements or attributes in the form of a schema which facilitates the recording of information. Traditionally metadata has been divided into three different types [1]:

> *Descriptive* metadata describes a resource for purposes such as discovery or identification and therefore includes elements such as name or title, author, identifier, subject or keywords.
>
> *Structural* metadata provides an indication of how compound resources are organised, for example how the component parts of a web page are arranged.
>
> *Administrative* metadata provides information necessary in managing a resource, such as who can access it or when and how the resource was created.

More recently, additional categories of metadata have been separated out due to the colossal amounts of electronic data now being produced [2]:

> *Technical* metadata includes hardware, software applications and file formats since many science and engineering data formats require specific configurations.
>
> *Use* metadata is increasing in importance with the expectation and anticipation that electronic resources and data will be repurposed and reused in order to maximise their potential.
>
> *Preservation* metadata is an essential part of all digital preservation strategies which aim to improve the longevity of electronic information.

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 activities should be viewed as an integral part of sound data management practice. As a result, metadata that supports curation and preservation should be embedded into the core metadata for managing crystallography data.

# 2. Enabling Use, Reuse and Repurposing

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 [3]. In addition, well managed and curated data has the potential to be re-purposed and generate new and innovative scientific results. Recapturing and reproducing some types of data is sometimes 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. In terms of the crystallography community, the long-term provision of data is particularly important since structure determination can only be truly repeated or verified when the raw data is available. In addition, the availability of raw data is extremely useful for reanalysis and reprocessing as improved methods for performing these tasks emerge.

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 [4]. 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, cultural, political, organisational, financial and legal as well as technical issues.

When considering existing digital data for reuse, a researcher is likely to contemplate many questions in relation to the data, including:

- Who created the data? and under what conditions?
- What format is it in?
- What is the intellectual property (IPR) associated with the data?
- What software is required to access and process the data?
- Is that software currently available?
- Has the data been modified since it was created?
- If so, who made the changes and why?
- Is the data what it claims to be?
- Is it related to any other resources? If so, how?
- Are there any dependencies between the resources?
- Is there a data dictionary to help interpret the semantics?

Metadata can be used to provide answers to all of these questions and more; it also facilitates subsequent management of the content of a repository and aids the processes of selection and appraisal in the preservation of ingested material. Digital materials require constant maintenance and migration to new formats as technology changes. In order to survive into the future, the resources need preservation metadata that can exist independently from the systems which were used to create them. Preservation metadata is information that supports and documents the digital preservation process. It is sometimes considered a subset of technical or administrative metadata and incorporates:

- Provenance: *Who has had custody or ownership of the digital object?*
- Authenticity: *Is the digital object what it purports to be?*
- Preservation Activity: *What has been done to preserve the digital object?*
- Technical Environment: *What is required to render and use the digital object?*
- Rights Management: *What intellectual property rights must be observed?*

## 2.1 Preservation Strategies

Strategies for improving the longevity of digital data so that it remains fit for both contemporary use as well as reuse in the future will vary depending on the data, its characteristics and dependencies. Preservation strategies for digital resources can be divided into three main types, technology preservation; technology emulation and information

migration. Technology preservation involves the maintenance of a digital resource together with all of the hardware and software needed to interpret it. There are issues with storage, on-going maintenance and missing documentation which can result in museums of ageing and incompatible computer hardware. Nonetheless, the technique does have a short-term role for supporting the rescue of digital resources, also known as *digital archaeology*.

Technology emulation is concerned with preserving the original bit-streams and application software. Emulator programs or virtual machines that mimic the behaviour of obsolete hardware and software are used to access the original bit-stream. This technique is already widely used for the preservation of computer games since it tends to retain the "look and feel". Technology emulation reduces the need for regular digital object transformations, although the emulators and virtual machines may themselves need to be migrated. This approach has the greatest potential for digital resources that are complex or dependent on executable code.

Information migration, typically involves the periodic transformation of digital data from one file format to another. It is likely to result in a modification of the original bit-stream and consequently can result in problems with ensuring the integrity and authenticity of data, making it essential to document any preservation actions undertaken as part of preservation metadata. This method is a widely used solution by data archives and software vendors (e.g. a linear migration strategy is used by software vendors for some types of data such as Microsoft Office files). The focus of this strategy is on preserving the intellectual content of digital resources. The Reference Model for an Open Archival System (OAIS) [5] identifies four different types of migration strategy: *refreshment; replication; repackaging* and *transformation.* Migration is often combined with some form of standardisation or normalisation process [6] so that on ingest to a repository all the data files are in one of a small number of formats which can then be managed more easily.

Preservation strategies are not mutually exclusive and it is likely that elements of different strategies will be chosen to work together [7]. A commonly accepted suggestion is that the original bits together with documentation should be kept in perpetuity. Whatever strategy is formulated for a specific set of data, there will be implications for the technical infrastructure as well as he corresponding metadata and rights management.

## 2.2 Metadata Creation & Capture

It is prudent to record metadata as early as the planning and design stages will allow, as well as throughout the lifecycle of the digital resource or data since it becomes time-consuming and uneconomical to start attaching metadata after the event. For example, if metadata created by a digital camera at recording time is not stored immediately, it may have to be restored afterwards, manually and with great effort. Therefore, it is necessary for the different groups of stakeholders in the lifecycle to cooperate using compatible methods and standards.

Metadata can be stored either internally, in the same file as the content data, or externally in a separate file or in a dedicated metadata registry. Typical examples of embedded metadata include: EXIF metadata within photographs, TIFF headers and file properties in Microsoft Office programs. Tools have been developed to extract some of this metadata automatically, for example the National Library of New Zealand preservation metadata extraction tool [8]; JHOVE (JSTOR/Harvard Object Validation Environment) [9] for the identification and validation of formats; and the DROID (Digital Record Object Identification) tool [10] developed by The National Archives. Other means of automatic metadata extraction include text-mining. However, the trustworthiness of automatically extracted metadata is not entirely reliable, so that in many cases labour-intensive checking and validation is required. Also, many of the tools are incapable of dealing with conflicting metadata.

Other types of metadata are maintained in files tightly associated with the resources they describe rather than being embedded within them (e.g. README files or documentation; Databases e.g. bibliographic catalogues and e-journal systems; Documentation standards or codebooks). Some metadata may be created as part of the ingest process into a repository or automatically captured from the ongoing management of resources, recording for example custodial history; format transformations and usage information.

Registries are often used to store metadata created by third parties, since they may not have direct control over or access to the content of the resource. Typically such content is made available for harvesting through the Open Archives Initiative — Protocol for Metadata Harvesting (OAI-PMH) [11].

In addition, there is the question of data format: storing metadata in a human-readable format such as XML can be useful because users can understand and edit it without special tools. On the other hand, these formats are not optimized for storage capacity; it may be useful to store metadata in a binary, non-human readable format instead to speed up transfer and save memory.

Metadata creation, capture and maintenance is generally regarded as being tedious, time-consuming, subjective and labour-intensive and therefore very costly [12]. However, given the role that metadata plays in the management and curation of electronic resources it is important to get the structure and architecture of the metadata correct from the outset, as well as the appropriateness of the information that is recorded.

## 2.3 Metadata Schema

Metadata is made up of a number of elements or attributes which can be categorised according to the different functions they support. A structured set of metadata elements is normally organised into a schema, representing a data model and the attributes associated with the entities within it. There are several ways of encoding metadata which are in common use. HTML encoded metadata accounts for the majority of metadata embedded within Web resources and therefore readily available for harvesting. This approach has the great virtue of fitting in with existing Web infrastructure which already provides for mark up and communication protocols (HTTP). XML mark up tends to be used for the encoding and exchange of structured data, a particular strength being machine-processibility. In addition, the XML namespace facility provides structural capabilities that HTML lacks, making it easier to achieve the principles of modularity and extensibility. The Resource Description Framework (RDF) is the primary enabling infrastructure of Semantic Web activity. RDF is an additional layer on top of XML that is intended to simplify the reuse of vocabulary terms across namespaces.

## 2.4 Metadata Application Profiles

Metadata vocabularies or schemas lie at the heart of any application; they determine the functionality that an application is capable of delivering in terms of the services and information that it can provide. An application specific metadata vocabulary consists of terms, their definitions and any constraints relevant to the application.

A single metadata element set cannot be expected to accommodate the functional requirements of all applications; far more useful in this respect is the notion of an Application Profile (AP) — a metadata schema which draws on existing metadata element sets, adapting and customising specific elements for a particular local application [13,14]. The purpose of an AP is to adapt or combine existing schemas into a package that is tailored to the functional

requirements of a particular application, whilst retaining interoperability with the original base schemas. This offers the potential for digital materials to be accessed, used and curated effectively both within and beyond the communities in which they were created.

The JISC has recently recognised the potential of APs as elements of shared infrastructure for the research and education communities. To this end, a scoping study to investigate metadata AP requirements for scientific data in relation to digital repositories, and specifically concerning descriptive metadata to support resource discovery as well as other functions such as preservation, has recently been commissioned [15]. This follows the development of the Scholarly Works Application Profile (SWAP) [16]. APs for images, time based media as well as geospatial data are also being investigated. Arguably, scientific data encompass a much wider range of resource types and are far more complex than other kinds of material, so the Scientific Data Application Profile (SDAP) study explores whether harmonisation around an AP to improve resource discovery and reuse of scientific and research data in the repository landscape can be achieved or is even desirable. One of the most important models in this context is the CCLRC Scientific Metadata Model [17,18], see section 2.5 below.

In addition, further important work on APs for research datasets has been undertaken by the DRYAD [19] and DISC-UK Datashare [20] projects. DRYAD is a digital data repository for datasets underlying publications in the field of evolutionary biology; the associated metadata AP includes elements from: Dublin Core, Darwin Core, PREMIS, DDI and EML. DISC-UK Datashare aims to support academics that wish to openly share datasets, presenting a model for depositing 'orphaned datasets', which are not being deposited in subject-domain data archives or centres. Outputs from the project are intended to assist repository managers in overcoming barriers to incorporating research datasets into institutional repositories. The metadata AP is largely drawn from the Dublin Core Metadata Element Set [21] for cross-domain resource discovery.

## 2.5 CCLRC Scientific Metadata Model

The Science and Technologies Facility Council (STFC) [22] formerly known as the Council for the Central Laboratory of the Research Councils (CCLRC) is based in the UK providing one of Europe's largest multidisciplinary research support organisations. Operating from three sites, the STFC provides access to several large scale scientific facilities, such as accelerators, lasers, telescopes, satellites and supercomputers, for the UK research and industrial communities. Collectively, these facilities generate copious amounts of data; a trend which is set to rise from Terabytes into the order of Petabytes with the development of new instruments and facilities such as the DIAMOND Light Source (DLS) [23] and the Large Hadron Collider based at CERN [24]. The data generated and stored at STFC spans most major scientific disciplines including, astronomy, biology, chemistry, environmental science and physics.

The CCLRC Scientific Metadata Model (CSMD) [17] was developed as a means of capturing sufficient information relating to scientific studies and the data that they produce to enable sharing and reuse of such data for parallel and follow-on studies. The Model is in production use at various STFC facilities, including the areas of Neutron Science, Lasers and Synchrotron Science. The aim of the CSMD is to provide a high-level generic metadata model which can be specialised and adapted to specific scientific disciplines.

### 2.5.1 Data Model

The CSMD attempts to capture scientific activity at several different levels, see figure 1. At the top most level there is a *Policy* which drives research by initiating and maintaining one or more *Programmes* of work which in turn comprise one or more generic activities in the form

of studies or projects based on a particular theme or topic. A *Study* is a piece of work performed by a principal investigator and/or institution, along with co-investigators and researchers. Since a *Study* is normally funded by a *Programme* (e.g. the UK e-Science Programme) it may also have a grant number associated with it.

A *Study* comprises *Investigations* which can be of different types and typically involve data collection stages. The model explicitly considers three types of *Investigation*, although it is possible to add others:

- An *Experiment* is an investigation into the physical behaviour of the environment usually to test a hypothesis. It typically consists of a controlled environment involving an instrument operating under constrained settings and environmental conditions.
- A *Measurement* is an investigation that records the state of some aspect of the environment at specified intervals of time and space, generally using a passive detector.
- A *Simulation* takes a mathematical model of a part of the world and from a set of initial parameters determines how the modelled system reacts or evolves over time.

The CSMD also recognizes the existence of *Virtual Studies*. These are groups of studies that are related in some way, perhaps having the same principal investigator or institution and subject matter, but funded under differing *Programmes.*

The collected data itself is covered in a separate model, although shown combined above in Figure 1. Each *Investigation* produces a *Data Holding*. This *Data Holding* is made up of one or more *Data Collections*, each of which may comprise further *Data Collections*. The concept of a *Data Collection* enables different sets of data to be separated out, e.g. raw instrument data from the intermediate and processed sets of data. Each *Data Collection* is ultimately represented by a set of *Atomic Data Objects*, which are physical data files or database queries from which the data may be obtained.

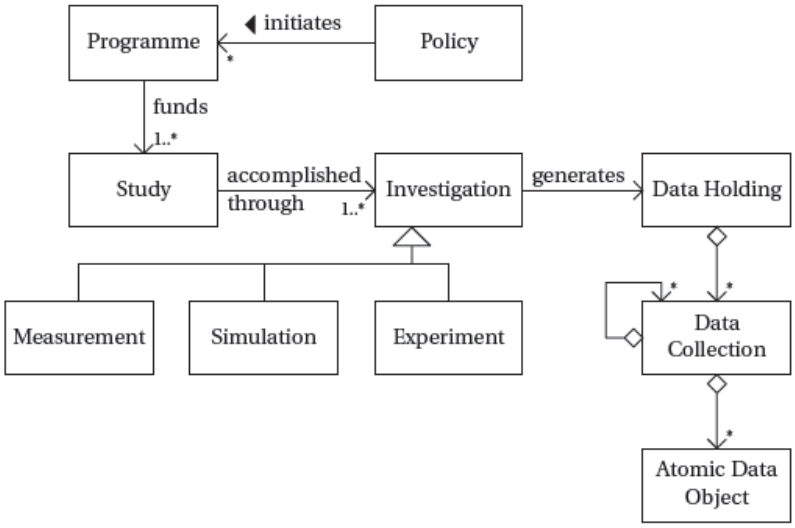


Figure 1: CSMD Data Model reproduced from the SDAP study [15]

In addition there is provision for recording supplementary information relating to a *Study* including the following:

- Topic Indexing (Keywords, Taxonomies)
- Provenance (What the study is, who did it and when)
- Data Holding
  - Detailed description about the data and its organisation

    - Parameters information
    - Atomic Data Objects can refer to regular files or database queries
    - Some workflow elements
- Legal notes, copyright, patents and conditions of use relating to the study and the data in the study
- Related material, publications, community information and associated links
- Access Conditions

### 2.5.2 ICAT

The storage, retrieval and management of data are a major concern for all large scale facilities. For example, ISIS [25], based at STFC and operating since 1984, currently produces ~1TB of neutron and muon data each year, with this rate of data collection set to rise still further. The full value of these data resources will only be realised if they are easily searchable, accessible and reusable. A simplified version of the CSMD is being used in the ICAT database which provides an interface to over 20 years of ISIS experiment data. ICAT is a database, with a well defined API, that provides a uniform interface to experimental data and a mechanism to link all aspects of research from proposal through to publication.

The ICAT database, together with supporting software is also being used to develop a web-based data portal (STFC DataPortal [26]) with the objective of offering a simple method for browsing and searching the contents of all STFC data resources. ICAT is also being integrated into the DIAMOND Light Source (DLS) and Central Laser Facilities (CLF) at STFC.

ICAT is further being used in a variety of applications external to the STFC and including DataMINX in Australia [27] as well as applications at the Institut Laue-Langevin (ILL) based in Grenoble, France.

# 3. Preservation Metadata Standards

The Open Archival Information System (OAIS) Reference Model [5] has been very 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 3.1); packaging information and descriptive information. *Representation Information* is defined as any information required to render, process, interpret, use and understand (digital) data. For example, it may be a technical specification, or a data dictionary or a software tool. Also, within an OAIS, information is encapsulated in packages comprising: content information, preservation description information and *packaging information*. Packaging information comprises data relating to one of the processes: submission; archival or dissemination.

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 several 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 [28], CEDARS project [29], NEDLIB Project [30]). However, in 2002, the OCLC/RLG Preservation Metadata Framework Working Group consolidated existing expertise in the form of a preservation metadata framework [31]. 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 firstly, to define a core set of implementable, broadly applicable preservation metadata elements, supported by a data dictionary; and secondly to 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 [32]. In May 2005, PREMIS followed up the survey report with the *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group* [33]. The report included: 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 was set up to manage the evolution of the Data Dictionary [34] which has resulted in the PREMIS Data Dictionary V2.0 [35], released in March 2008.

## 3.1 OAIS Preservation Description Information

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 in 2003 (ISO 14721:2003 [5]). The word "Open" in the title refers to the mechanism used in the development of the model (i.e. within an open forum) rather than to the open availability of the content in an archive — it is therefore equally applicable to dark as well as open archives. The model has recently undergone an open review process and a revision is imminent.

*Preservation Description Information (PDI)* is defined as information or metadata "*which will allow the understanding of the Content Information over an indefinite period of time*". The standard describes PDI 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 with 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).

## 3.2 PREMIS Data Dictionary V2.0

Although the OAIS Reference Model and its concepts of representation information, packaging information and PDI remain 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 [36]. The PREMIS data dictionary provides an intermediate stage in between OAIS PDI and an actual application

specific implementation; its major functions are to cater for data exchange and interoperability.

The PREMIS group defines preservation metadata as, "*the information a repository uses to support the digital preservation process*". In particular, the group looked at metadata supporting the functions of maintaining *viability*, *renderability*, *understandability*, *authenticity,* and *identity* in a preservation context. Although, these terms are not defined by the PREMIS group, Priscilla Caplan has elaborated on them in a recent Library Technology Report [36].

- *Viability* is the quality of being readable from media, with media deterioration and media obsolescence being the main threats.
- *Renderability* involves ensuring that a digital file is displayable, playable or otherwise usable as appropriate.
- *Understandability* requires that enough information in the form of metadata, documentation and/or related objects should be maintained to enable interpretation and understanding by future users.
- *Authenticity* is often defined as "*the quality that an object is what it purports to be*". Preservation treatment should not compromise data integrity, i.e. ensure that the object is not destroyed or modified in an unauthorised manner.
- *Identity* – persistent identifiers should be used to reference preservation objects to ensure their long-term access.

Preservation metadata thus cross-cuts a number of the categories typically used to differentiate the various types of metadata: administrative (including rights and permissions), technical, and structural. Particular attention was paid to the documentation of digital provenance (the custodial history of an object) and to the documentation of relationships, especially relationships among different objects within the preservation repository.

In establishing a core set of preservation metadata, the PREMIS group uses a practical definition i.e. "*the information most preservation repositories need to know to preserve digital materials over the long-term*". Note, however that "core" does not necessarily mean mandatory. Ease of implementation was also one of the guiding principles of the data dictionary, which places emphasis on rigorous definitions supported by usage guidelines and recommendations with a particular focus on automated metadata collection.

Version 2.0 of the PREMIS Data Dictionary was released in March 2008[35]. Major changes in this revision include:

- Expanded rights metadata; differentiation among several types of intellectual property rights, including copyright, statue and license
- More extensive significant properties and preservation level information
- A mechanism for extensibility for a number of metadata units
- Explicit provision for referencing file format registries and digital signatures

### 3.2.1 Data Model

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). In the first version of the Data Dictionary, relationships between Rights and Agents and between Events and Agents were defined as unidirectional; in version 2.0 all relationships are defined as bi-directional.

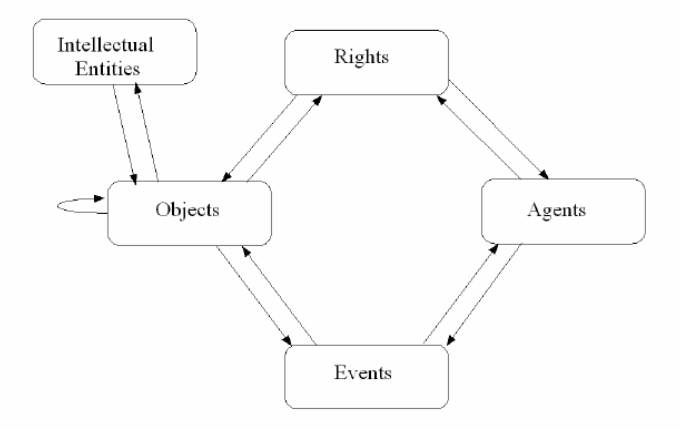


Figure 2: PREMIS Data Dictionary Data Model V2.0 [35]

*An Intellectual Entity* is a set of content that is considered a single intellectual unit for purposes of management and description, 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 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. A file can be zero or more bytes and has a file format, access permissions, and file system characteristics such as size and last modification date.
>
> REPRESENTATION: the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.
>
> BITSTREAM: is contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes. A bit stream cannot be transformed into a standalone file without the addition of file structure (headers, etc.) and/or reformatting the bit stream to comply with some particular file format.

*An Event* is an action that involves or impacts at least one Object or Agent associated with or known by the preservation repository.

*An Agent* can be a person, organization, or software program/system associated with Events in the life of an Object, or with Rights attached to 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. Note that PREMIS deals only with rights and permissions related to preservation activities, leaving aside those concerned with access and dissemination.

### 3.2.2 Semantic Units

The PREMIS group has defined *Semantic Units* rather than "metadata elements". Each semantic unit defined in the Data Dictionary is mapped to one of the entities in the data model. In this sense, a semantic unit may be viewed as a property of an entity. For example, the semantic unit *size* is a property of an Object entity.

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

In terms of mapping between PREMIS Semantic Units and OAIS PDI:

- *Provenance* information corresponds to the semantic units associated with the *Events* entity
- *Reference* information corresponds to the **objectIdentifier** semantic unit
- *Context* information is covered by **linkingEventIdentifier,** l**inkingIntellectualEntityIdentifier**, **linkingPermissionStatementIdentifier**
- *Fixity* information is covered by the **signatureInformation** and **fixity** semantic units

# 5. Crystallography Data

Critical to developing appropriate and adequate metadata is a thorough understanding of the nature and characteristics of the data as well as the workflows and processes involved in generating it. It is, however, clear that processes and workflows in each crystallography laboratory differ considerably [38,39]. A key requirement is an understanding of the file formats in use as well as the inter-relationships between processing software and data files.

## 5.1 Background

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 [40].

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 [41]. 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 [42], 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.

## 5.2 An Exemplar Repository: eCrystals@Soton

eCrystals@Soton [43] 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 [44]. 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.

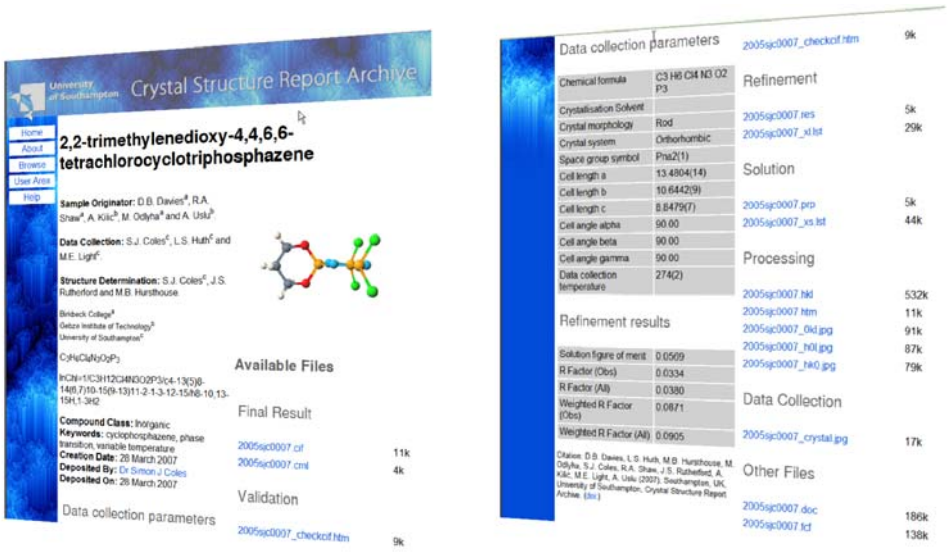


Figure 3: An example Crystal Structure Report in eCrystals [43].

An analysis of the work processes at the NCS 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.

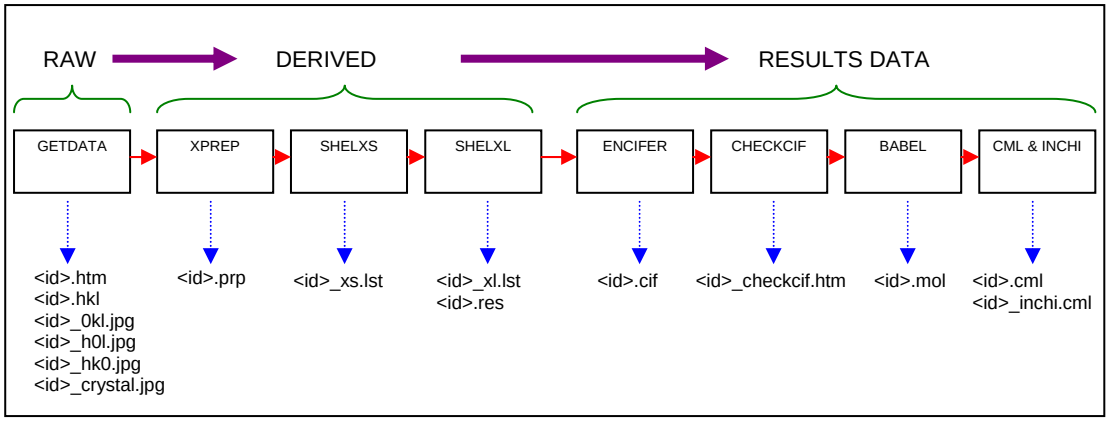


Figure 4: Workflow model of the EPSRC UK National Crystallography Service (NCS)

The primary aim of the eCrystals repository is to make available and encourage the sharing of data, which is generated throughout the experiment pipeline. The screen shot in Figure 3 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 [41] 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, the format 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 [45]. 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 [46], 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 [47] 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 [48] 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 3, a 3D graphical rendition of the molecule (jmol) is also generated and can be rotated interactively on the eCrystals website.

## 5.3 The eBank-UK Application Profile

In establishing the eCrystals repository, the eBank-UK project developed a metadata application profile which was designed to help manage crystallography data and address several objectives, including: resource discovery and use; data interoperability and exchange; the linking of publications and their underlying datasets; automatic and semi-automatic metadata generation; data and metadata quality control and data security [49].

The eBank-UK AP is largely based on Dublin Core metadata and can be considered to be a Dublin Core Application Profile providing for consistency, long-term quality control and interoperability with other metadata schema. It is generally accepted that the 15 elements of the Dublin Core Metadata Element Set (ISO 15836:2003E) form the minimum information required for adequate description, administration and technical management of a digital resource. Furthermore, they are the metadata elements which are most likely to be used for creating indexes and forming search queries. These same elements are also required to enable metadata harvesting using the OAI-PMH [11] as well as federated searching.

At present, eCrystals uses the eBank-UK Metadata Application Profile [49] from which much of the necessary curation and preservation metadata is absent since preservation functionality was not a priority when the repository was first constructed. A full and comprehensive

exposition of the development of the eBank-UK AP [49] is available from the eBank-UK project website. The AP is currently encoded in the XML schema language (XSD). 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*

Reference information: the eCrystals data repository currently uses Digital Object Identifiers [50] as a form of reference identifier as well as the IUPAC International Chemical Identifier (InChi) [46] as a domain identifier.

Provenance information: versioning is the only type of information currently stored by the ePrints.org software upon which eCrystals has been built.

Context information: the only type of relationship recorded in ePrints.org at present is that of versioning information; however, other types of context information should be recorded according to the adopted preservation strategy.

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.

Rights Information (copyright, IPR, preservation rights): a rights and citation statement is available on the eCrystals website [43].

## 5.4 Proposals for Preservation Metadata

The primary aim of preservation metadata is to support preservation activities; consequently, differing preservation strategies are likely to demand that distinct types of information be recorded. For example, a preservation plan based on migration activities will require different information to that of one based on emulation. Hence, the preservation plans and policies of a particular repository will heavily influence the specific metadata that is to be recorded.

Metadata plays a crucial rule in ensuring that high quality documentation and community knowledge associated with a particular set of data are properly captured and made available across the entire lifecycle of the data; from the early stages of an experiment to secondary analysis by other researchers or use by policy makers and other key stakeholders. Metadata is used to facilitate understanding, reuse and management of data; however, the metadata required for effective data management varies with the type of data and its context of use.

The DCC Curation Lifecycle Model [51] provides a graphical, high level overview of the stages required for successful curation and preservation of data from its initial conceptualisation to its life end. 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; consequently there is a need to capture metadata at each stage of the cycle. 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)*

Whilst the exact metadata to be recorded is dependent on the individual preservation strategy in force, there is some consensus on a core set of preservation metadata as exemplified by the PREMIS Data Dictionary [35]. 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 [52] which examines the take-up of the PREMIS Data Dictionary and the implementation issues that have been encountered by 16 repositories. In addition, there is *Preservation Metadata for Institutional Repositories: applying PREMIS* [53] and *A Review of metadata standards in use by SHERPA DP repositories* [54].

According to Priscilla Caplan [55], 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 is necessary to use file format registries such as PRONOM [56] or the GDFR [57]. For automated extraction of format information, tools such as JHOVE [9] and DROID [10] can be used. There is also a need to take account of the standards and formats adopted within the user community.

**Significant Properties:** these are characteristics which should be retained throughout future preservation activities [58], in order to maintain understanding and renderability.

**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 (i.e. events and processes performed on the object).

Before the publication of the PREMIS Data Dictionary V2.0, the original version of the Data Dictionary was taken as a starting point in attempting to identify Semantic Units that may be of relevance for crystallography data and resulted in the following suggestions:

**Object Entity:**

objectIdentifer, preservationLevel, objectCharacterisitcs, creatingApplication, storage, environment, relationship, linkingEventIdentifier, linkingIntellectualEntityIdentifier, linkingPermissionStatementIdentifier

**Event Entity**
eventIdentifier, eventType, eventDateTime, eventDetail, eventOutcomeInformation, linkingAgentIdentifier, linkingObjectIdentifer

**Agent Entity**
agentIdentifier

**Rights Entity**
permissionStatementIdentifier, linkingObject, permissionGranted, grantingAgent

Within the eCrystals Federation project, we began by contemplating ways of supplementing the metadata in the eBank-UK AP with preservation metadata for crystallography data. This could be achieved either by revision or extension of the current eBank-UK Metadata AP, or by implementing a separate eBank Preservation Metadata Profile; the appropriateness of these strategies need to assessed in the context of the eCrystals Federation as well as the crystallography and Chemistry communities. This work was however superseded by developments within the crystallography community which resulted in the setting up of the Crystallography Data Commons and a decision was taken by the project to contribute to the development of a schema capable of addressing all the experimental data from a crystallography structure determination workflow rather than to extend the eBank-UK AP.

# 6. Crystallography Data Commons

A collaborative venture named *Towards an International Data Commons for Crystallography* (TIDCC) emerged as a result of discussions between participants in the TARDIS (The Australian Repositories for Diffraction Images), eCrystals Federation and DataMINX Projects and the Australian Research Council's Molecular & Materials Structure Network (MMSN) in September 2008.

TARDIS [59] is a multi-institutional collaboration that aims to facilitate the archiving and sharing of raw X-ray diffraction images from the protein crystallography community. Whereas the model coordinates and less often the structure factors (processed experimental data) are stored in the Protein Data Bank [60] the raw diffraction data is rarely made available. Consequently, there is a pressing need for the archiving and curation of raw X-ray diffraction data. However, the relatively large size of these datasets has presented challenges for storage in a single worldwide repository. This problem can be avoided by using a federated approach, where each institution or university utilizes its institutional repository.

The Australian National Collaborative Research Infrastructure Strategy (NCRIS) is a Federal Government initiative to renew and enhance Australian research infrastructure. DataMINX [61] is an NCRIS supported project to deliver data access, transport, management and repository services for the research community using and operating microscopy imaging, neutron and X-ray facilities for the determination of molecular and materials structure.

The MMSN [62] links scientists, technicians and students engaged in the determination and analysis of atomic structures of any kind; biological molecules, chemical molecules or solid state materials – and unites them with Grid computing, visualisation, database, informatics and applied mathematics researchers.

The intention of the TIDCC is to develop a community derived metadata schema capable of describing all types of crystallography data related to single crystal diffraction and that would loosely map onto the CSMD [17] (see section 2.5) by initially merging the TARDIS and eBank-UK schemas. The eventual aim being to create a METS profile [63], which would be used as a basis for the metadata platform on top of which could be built protocol based

applications using for example SWORD [64], ATOM [65], OAI-ORE [66] etc. for publishing, managing and preserving crystallographic data.

While we suggest a set of core preservation metadata below, it should be noted that such a set of information would need to be extended to include further and more detailed metadata to support a particular preservation strategy and the activities associated with it. Also, it should be stressed that the work described below is still very much in progress and in a constant state of flux.

## 6.1 Draft TIDCC Data Model

As a step towards developing a profile for federated repositories of crystallography data, the collaborators proposed a data model based on a typical crystal structure determination workflow — that of the EPSRC NCS based at the University of Southampton (see section 5.2). Figure 5 shows the proposed data model on which the TIDCC Metadata Application Profile (TMAP) is based.

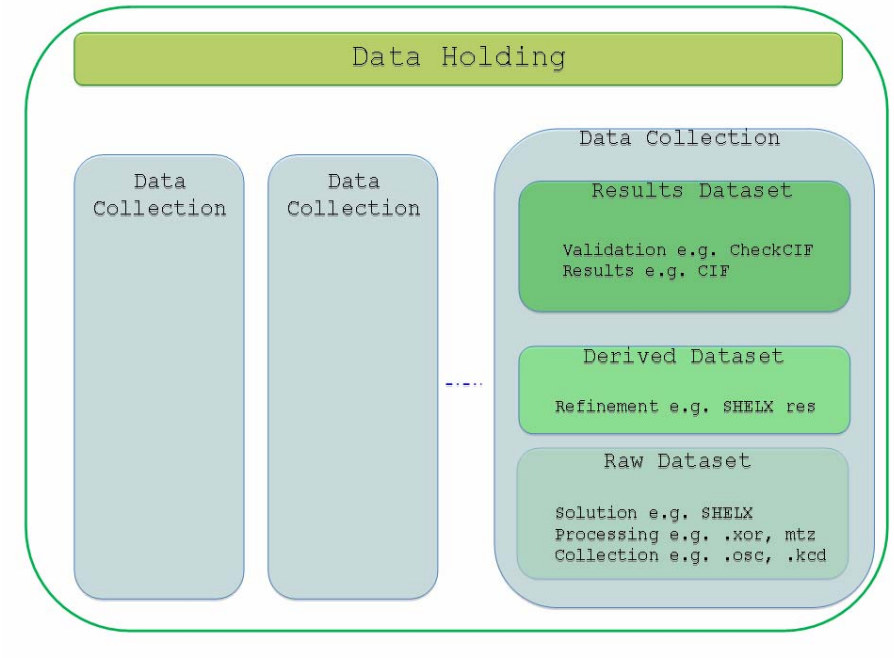


Figure 5: Draft TIDCC Data Model based on the workflow of the EPSRC NCS

The Model maps onto the lower dimensions of the CSMD (see Figure 1) albeit with slightly differing concepts:

TIDCC: Data Holding corresponds to CSMD: Study
TIDCC: Data Collection corresponds to CSMD: Data Holding/Investigation
TIDCC: Dataset corresponds to CSMD: Data Collection
TIDCC: Data File corresponds to CSMD: Atomic Data Object

## 6.2 Draft TIDCC Metadata Application Profile

The objective of the TMAP is to layer metadata into several levels to facilitate conformance by a wide variety of repositories and applications. In addition, the TMAP aims to cater for publication (resource discovery) and dissemination of crystallography datasets as well as their

management and reuse. Consequently, the TMAP is structured to reflect the following categories of metadata:

> **Generic Description**: this is the minimum information required to describe a resource independently.
>
> **Sub-domain context**: information to enable automated third party services to determine deeper sub-discipline specific context.
>
> **Lab (Private) management**: experiment and sample details to enable discovery and reprocessing in a laboratory.
>
> **Preservation**: all the information necessary to enable a third party to inherit and/or manage data at any point in the future.

The proposed architecture and metadata elements of the TMAP are presented in an Appendix to this report, but note that this work is still evolving and very much in a state of continuous change. The TMAP draws together relevant schema, including representations from the eBank-UK AP; the TARDIS schema; the CSMD and associated schema [17] and the ICAT schema from ICAT V3 Data Dictionary (2009) [67]. Once again, it should be borne in mind that the proposed profile is generic in nature and would need to be supplemented with additional metadata depending on the preservation policy and strategy adopted by a particular repository.

# 7. Community & Federation Issues

As mentioned in the introduction, the technical aspects of digital curation and preservation are only one facet of a multidimensional problem; curatorial issues further encompass social, cultural, political, organisational, financial and legal factors as well. Community consensus and the development of standards and guidelines for best practice underpin the longevity, effective management, preservation, sharing and reuse of science data.

## 7.1 Community Consensus

There is no doubt that appropriate metadata is crucial to the long-term availability and reuse of data. However, for such data to be shared, exchanged, maintained and used by third parties it is necessary to achieve agreements regarding the metadata vocabulary to be used as well as community consensus with regard to the definitions and constraints on specific metadata elements.

Published metadata specifications are often held in a central location, such as a reference document on a website or in a web accessible metadata registry. They generally contain semantic definitions of the elements and standard ways of representing them in digital formats such as databases and XML or RDF, which are rapidly becoming the de facto mark-up standards in many communities. Semantic definitions include both *Metadata Structure Standards* and *Metadata Content Standards*. The former ensure consistent structure to enable data sharing and searching, manage the creation process, record provenance and technical processes and manage access permissions, while the latter ensure effective machine searches through consistent data entry and the inclusion of access points using controlled vocabularies such as authority files, thesauri or encoding schemes.

In general, the process of designing a new metadata element set involves a great deal of effort and can be very time-consuming, particularly where widespread consensus is a necessity. Moreover, the use of new vocabularies has a tendency to undermine interoperability between disparate services and applications. The reuse of terms from existing vocabularies promotes

convergence or harmonisation within specific domains or application areas, and is an important step towards interoperation of a diverse range of applications.

The TMAP (see section 6) was a first attempt at constructing an over-arching AP for crystallography data which would facilitate the exchange of not only metadata, but also the data itself. However, following several meetings is has become apparent that a more effective way forward is to adapt the ICAT data model and schema to cater for curatorial and preservation activities since ICAT is presently being used by a growing proportion of the science community for managing their data (see section 2.5). This work is still in progress; it is expected that the preservation metadata proposed in the TMAP will feed into this development.

## 7.2 Shared Infrastructure

Shared infrastructure is important in terms of interoperability, not only to support the exchange of data between repositories, but also as a means of reusing existing metadata. Both the OAI-PMH [11] and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [66] have an important role to play in this respect.

Community agreement on a crystallography metadata application profile is likely to lead to improvements in the quality and consistency of both the metadata and the scientific data that is made openly available for sharing and reuse. This will in addition provide support for the aggregation of datasets into larger collections that could potentially support complex searches and mining operations. Moreover, the likelihood of publication and dissemination will promote the integration of metadata capture at all stages of the lifecycle, so that events are documented as they occur for curation and preservation purposes.

## 7.3 Curation & Preservation

Although a first set of proposed preservation metadata for crystallography data has been put forward as part of TMAP, there is still some work remaining in choosing appropriate Semantic Units from the PREMIS Data Dictionary V2.0 and incorporating them into the ICAT schema.

# 8. Conclusions

It is clear that the crystallography community recognises the importance of high quality metadata for all the functions that it can support, including the long-term accessibility and reuse of scientific data. Data which is better documented is easier to find and use, and is likely to be of greater consistency and quality, enhancing the ability to replicate and validate experimental results using the actual data and processes used in the original investigation or study. Reliable metadata also promotes the development of improved tools for data management, to assist data producers, librarians, and archivists as well as the establishment of virtual research communities. With well packaged information which combines research findings, data and metadata, it becomes possible to maintain linkages between primary and secondary datasets and publications, providing for richer comparison and broader knowledge.

Although there is still a considerable way to go along the path to formulating community agreed metadata for the curation and preservation of crystallography data, the work outlined in this report proves that the crystallography community appreciates the benefits and does not lack the motivation to achieve such as goal.

# References


1. Murtha Baca (Ed.), *Introduction to Metadata: Pathways to Digital Information*, Getty Information Institute, 1998
2. Lord, P., Macdonald, A.: 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
3. 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/
4. Beagrie, N.: Digital Curation for Science, Digital Libraries, and Individuals, International Journal of Digital Curation, Vol. 1 (2006), http://www.ijdc.net/ijdc/article/view/6
5. Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System, ISO:14721:2002, 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf#search=%22OAIS%20model%22
6. Clifford Lynch, *Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information*, D-Lib Magazine, September 1999 5(9)
7. Manjula Patel, *Preservation Planning for Crystallography Data*, WP4, eCrystals Federation Project, 25th June 2009
8. National Library of New Zealand Metadata Extraction Tool, http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool
9. JHOVE - JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/
10. DROID (Digital Record Object Identification), The National Archives, http://droid.sourceforge.net/wiki/index.php/Introduction
11. Open Archives Initiative — Protocol for Metadata Harvesting (OAI-PMH) http://www.openarchives.org/OAI/openarchivesprotocol.html
12. 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
13. Rachel Heery & Manjula Patel, *Application Profiles: Mixing and matching metadata schemas*, ARIADNE Issue 25, September 2000
14. T. Baker, M. Dekkers, R. Heery, M. Patel, G. Salokhe, *What terms does your metadata use? Application profiles as machine-understandable narratives*, Journal of Digitial Information, Vol 2. Issue 2. November 2001
15. Alexander Ball, *Scientific Data Application Profile Scoping Study Report*, 2009, UKOLN, University of Bath, http://homes.ukoln.ac.uk/~ab318/docs/ball2009sda/
16. Julie Allinson, Pete Johnston and Andy Powell, *A Dublin Core Application Profile for Scholarly Works*, Ariadne Issue 50, January 2007, http://www.ariadne.ac.uk/issue50/allinson-et-al/
17. S Sufi & BMatthews, *CCLRC Scientific Metadata Model: Version 2*, Technical Report DL-TR-2004-001, 2004, CCLRC Daresbury Laboratory, Warrington, http://epubs.cclrc.ac.uk/work-details?w=30324
18. Shoaib Sufi and Brian Matthews, *A Metadata Model for the Discovery and Exploitation of Scientific Studies*, Knowledge and Data Management in Grids, Computer Science, Springer US, Feb 2007, http://www.springerlink.com/content/p758572242675671/fulltext.pdf
19. S. Carrier, *The Dryad Repository Application Profile: Process, Development, and Refinement,* 2008, DOI: http://hdl.handle.net/1901/534.
20. DISC-UK Datashare Project, http://www.disc-uk.org/index.html

21. Dublin Core Metadata Element Set, http://dublincore.org/documents/dces/
22. Science and Technologies Facility Council (STFC), http://www.scitech.ac.uk/
23. DIAMOND Light Source, STFC, http://www.scitech.ac.uk/PandS/Gallery/Diamondls.aspx
24. LHC –The Large Hadron Collider, CERN, http://lhc.web.cern.ch/lhc/
25. ISIS, STFC, http://www.scitech.ac.uk/About/wwd/ISIS.aspx
26. STFC DataPortal,
27. DataMINX Project, http://www.dataminx.org/
28. National Library of Australia, *Preservation Metadata for Digital Collections*, October 1999, http://www.nla.gov.au/preserve/pmeta.html
29. Cedars Project: *Guide to Preservation Metadata, http://www.leeds.ac.uk/cedars/guideto/metadata/*
30. Catherine Lupovici and Julien Masanès, *Metadata for Long-Term Preservation*, NEDLIB project, July 2000, http://nedlib.kb.nl/results/D4.2/D4.2.htm
31. 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
32. *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
33. *Data Dictionary for Preservation Metadata*, Final report of the PREMIS Working Group, May 2005, http://www.oclc.org/research/projects/pmwg/premis-final.pdf
34. PREMIS Preservation Metadata: Maintenance Activity, http://www.loc.gov/standards/premis/
35. PREMIS Data Dictionary for Preservation Metadata V2.0, March 2008, http://www.loc.gov/standards/premis/v2/premis-2-0.pdf
36. *Metadata Standards Framework –Preservation Metadata (revised)*, National Library of New Zealand, June 2003, http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised
37. Priscilla Caplan, *The Preservation of Digital Materials*, Library Technology Reports, Vol. 44 (2), February/March 2008
38. 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
39. *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
40. Allen, F. H., *High-throughput crystallography: the challenge of publishing, storing and using the results*. Crystallography Reviews, 10, pp3-15 (2004).
41. CIF -The Crystallographic Information File, http://www.iucr.org/iucr-top/cif/
42. CrystalEye, Unilever Centre for Molecular Informatics, University of Cambridge, http://wwmm.ch.cam.ac.uk/crystaleye/
43. The Crystal Structure Report Archive –eCrystals Data Repository, http://ecrystals.chem.soton.ac.uk
44. 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
45. Chemical Markup Langauage (CML), http://www.ch.ic.ac.uk/rzepa/cml/
46. IUPAC International Chemical Identifier (InChi), http://www.iupac.org/inchi/
47. Open Babel: The Open Source Chemistry Toolbox, http://openbabel.org/wiki/Main_Page
48. SHELX Home Page, http://shelx.uni-ac.gwdg.de/SHELX/

49. eBank-UK: Metadata Schemas, http://www.ukoln.ac.uk/projects/ebank-uk/schemas/
50. Digital Object Identifier System, DOI Foundation, http://www.doi.org/
51. 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
52. 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
53. 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
54. *A Review of metadata standards in use by SHERPA DP repositories*
55. Priscilla Caplan, *Instalment on "Preservation Metadata"*, DCC Curation Manual, July 2006, http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/
56. The PRONOM registry, The National Archives, http://www.nationalarchives.gov.uk/pronom/
57. The Global Digital Format Registry (GDFR), Digital Library Federation, http://hul.harvard.edu/gdfr/
58. Andrew Wilson, *Significant Properties Report*, InSPECT project, April 2007. http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf
59. The Australian Repositories for Diffraction Images – TARDIS Project, http://tardis.edu.au/
60. Protein Data Bank, http://www.rcsb.org/pdb/home/home.do
61. DataMINX Project, http://www.dataminx.org/
62. Australian Research Council's Molecular & Materials Structure Network (MMSN), http://mmsn.net.au/index.htm
63. Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets/
64. Julie Allinson, Sebastien François and Stuart Lewis, *SWORD: Simple Web-service Offering Repository Deposit*, Ariadne Issue 54, January 2008, http://www.ariadne.ac.uk/issue54/allinson-et-al/
65. The ATOM Publishing Protocol, RFC5023, October 2007, http://www.ietf.org/rfc/rfc5023.txt
66. Open Archives Initiative Object Reuse and Exchange (OAI-ORE), http://www.openarchives.org/ore/
67. ICAT, STFC, http://code.google.com/p/icatproject/

# Appendix: Draft TIDCC Metadata Application Profile

## Generic Description

| | | |
|---|---|---|
| Crystallography Data Holding | dc:type | mods:genre (with type attribute) |
| Chemical / Biological Name | dc:title | mods:title (subelement of mods:titleInfo) |
| Authors | dc:creator | mods:name |
| Author Affiliation | ?dc:publisher <XSD:element name ="InstitutionAffiliatedTo" | mods:affiliation (subelement of mods:name) |
| Organisaion making data available | dc:publisher | mods:publisher (subelement of mods:originInfo) |
| Description | dc:description | |
| Relationship between datafiles | dc:relation | |
| Creative Commons License / Rights? | dc:rights | mods:accessCondition |
| Resource Identifier | dc:identifier | mods:identifier (with type attribute) |
| Date of deposit | dc:created | |
| Domain Identifier (InChI / LSID) | http://purl.org/ebank/terms/InChI | As for keyword with different authority, else identifier type="domain" authority="http://purl.org/ebank/terms/InChI" |
| Resource Type | dc:type <xsd: element name: investigation type> <xsd: element name: dataholdingtype> | mods:genre (with type attribute) |

## Sub-Domain Context

| | |
|---|---|
| Version | dc:modified One of the mods:originInfo fields; mods:identifier type=”version_identfication”; or mods:note type=”version_identification” |
| Sub category (fine filter, e.g. Keywords) | http://purl.org/ebank/terms/Keywords<br>mods:subject authority=http://purl.org.ebank.terms/Keywords<br>or mods:subject authority=”iucr” then mods:topic or other mods:subject sub-element |
| Category (Coarse filter, Compound Class, etc) | http://purl.org/ebank/terms/CompoundClass<br>As per Keywords with different authority, else genre type=”category” authority=”http://purl.org/ebank/terms/CompoundClass” |
| (Chemical) Formula | http://purl.org/ebank/terms/ChemicalFormula<br>As for keyword with different authority, else identifier type=”formula” |
| Persistent Identifier | http://purl.org/ebank/terms/DOI mods:identifier type=”doi” |
| Free text box (Open Comment) | <xsd:simpleType="CommunityInformation"> |
| Constituent dataset types [Initialisation, Collection, Processing, Solution, Refinement, CIF/Result, Validation] | http://purl.org/ebank/terms/EbankDatasetType<br>?dc:IsPartOf? Add <xsd:element name="images"<br>from TARDIS mods:genre with type attribute<br><xsd:complexType name="DataDescriptionType"><br>element name dataname<br>element name typeofdata<br>element name softwaretype |

## Lab (Private) Management

| | |
|---|---|
| Date Experiment Performed | <xs:element name="date" substitutionGroup="dc:date"/><br><xsd:complexType name ="DateTimeType"> |
| Local Code | <xsd:element name="datasetName"<br><xsd:element name="crystalName" |
| Cell | length_a |
| | length_b |
| | length_c |
| | angle_alpha |
| | angle_beta |
| | angle_gamma |
| | volume |
| | formula_units_Z |
| Symmetry | space_group_name<br>cell_setting |
| Local Code | <xsd:element name="datasetName"<br><xsd:element name="crystalName" |
| Collection Temperature | cell_measurement_temperature<br>diffrn_ambient_temperature |
| Diffractometer Type | <xsd:element name="diffractometerType" |
| Data Collection Software | computing_data_collection <xsd:complexType name="softwareType"> production |
| Detector Type | <xsd:element name="detectorSN" |
| Image Format | <xsd:element name="imageType" ICAT3:Datafile_format |
| Wavelength | <xsd:element name="xraySource"<br><xsd:element name="xrayWavelength" |
| Exposure Time | <xsd:element name="exposureTime" |

| | |
|---|---|
| Resolution Limit | <xsd:element name="resolutionLimit" |
| Processing Software | computing_data_reduction <xsd:complexType name="softwareType"> |
| Solution Software | computing_structure_solution |
| Refinement Software | computing_structure_refinement |
| Agreement Statistics | refine_ls_number_parameters |
| | refine_ls_number_restraints |
| | refine_ls_R_factor_all |
| | refine_ls_R_factor_gt |
| | refine_ls_wR_factor_ref |
| | refine_ls_wR_factor_gt |
| | refine_ls_goodness_of_fit_ref |
| | refine_ls_restrained_S_all |
| | refine_ls_shift/su_max |
| | refine_ls_shift/su_mean |
| Morphology | exptl_crystal_description<br>exptl_crystal_colour |
| Oscillation Angle | <xsd:element name="oscillationAngle" |
| Oscillation Range | <xsd:element name="oscillationRange" |
| Oscillation Start | <xsd:element name="start" |
| Oscillation End | <xsd:element name="end" |
| Free text box (private comments) | ? |
| Date Experiment Performed | <xs:element name="date" substitutionGroup="dc:date"/><br><xsd:complexType name ="DateTimeType"> |

## Preservation

| | |
|---|---|
| Record Version | <xs:element name="isVersionOf" substitutionGroup="relation"/><br><xs:element name="replaces" substitutionGroup="relation"/> |
| Policy | ? |
| Publishing Institution | dc:publisher (mods:publisher (subelement of mods:originInfo) |
| Software Version | <xsd:complexType name="ProgramType"> element name version |
| Software Name | <xsd:complexType name="ProgramType"> element name ProgramName |
| Software Author / Origin | <xsd:complexType name="ProgramType"> element name uri |
| Final Checksum | ICAT3:Datafile_checksum |
| File Formats | <xs:element name="hasFormat" substitutionGroup="relation"/> |
| Provenence | ? |
| Enviromental Information | ? |
| | |

**Split Preservation into Raw data / derived data / Results data**

| | |
|---|---|
| **Results Data** | |
| Software Version | <xsd:complexType name="ProgramType"> element name version |
| Software Name | <xsd:complexType name="ProgramType"> element name ProgramName |
| Software Author / Origin | <xsd:complexType name="ProgramType"> element name uri |
| Final Checksum | ICAT3:Datafile_checksum |
| Final Checksum | ICAT3:Datafile_checksum |
| File Formats | <xs:element name="hasFormat" substitutionGroup="relation"/> |

| | |
|---|---|
| | |
| **Derived Data** | |
| Software Version | <xsd:complexType name="ProgramType"> element name version |
| Software Name | <xsd:complexType name="ProgramType"> element name ProgramName |
| Software Author / Origin | <xsd:complexType name="ProgramType"> element name uri |
| Final Checksum | ICAT3:Datafile_checksum |
| File Formats | <xs:element name="hasFormat" substitutionGroup="relation"/> |
| | |
| | |
| **Raw Data** | |
| Instumentatioin/ Manufacturer | <xsd:complexType name="ProgramType"> element name uri |
| Software Name | <xsd:complexType name="ProgramType"> element name ProgramName |
| Raw Data File Format | <xs:element name="hasFormat" substitutionGroup="relation"/> |
| Software version required to read raw images | <xsd:complexType name="ProgramType"> element name version |
| Final Checksum | ICAT3:Datafile_checksum |