
Representation Information for Crystallography Data

JISC eCrystals Federation Project

WP4: Repositories, Preservation and
Sustainability

Document Details

Author:	Manjula Patel (UKOLN & DCC)
Date:	19 th May 2009
Version:	1.2
File Name:	eCrystals-WP4-RI-090519.doc
Notes:	Final



This work is licensed under a [Creative Commons Attribution-Non-Commercial-Share Alike 2.5 UK: Scotland Licence](https://creativecommons.org/licenses/by-nc-sa/2.5/uk/).

Executive Summary

The eCrystals Federation project is concerned with setting up a federation of institutional repositories for the management and dissemination of the raw, derived and results data from crystallographic experiments [1]. It builds on the work of the eBank-UK project [2] which developed and implemented the eCrystals repository [3], focusing on the workflows of the laboratory based experimental technique of chemical crystallography undertaken at the EPSRC National Crystallography Centre (NCS) based in Southampton. Following the creation of a completed crystal structure determination, data is uploaded into eCrystals and supplemented with chemical and bibliographic metadata. A subsequent scoping study, undertaken as part of phase 3 of the eBank-UK project identified several issues pertinent to the curation and preservation of crystallography data [4], amongst them was the importance of the concepts underlying the OAIS Reference Model [5] and its associated notion of *Representation Information* (RI).

Consequently, this report is concerned with an investigation of RI for crystallography data and its role in the curation, maintenance and management of such data. We begin with a brief overview of aspects of the OAIS Reference Model [5] which establishes a conceptual framework of terms and components for use in the preservation of information. The Model also identifies the environment within which an OAIS operates as well as its basic functions in the form of functional, information and information flow models. The notions of a *Designated Community* (DC) and its associated *Knowledge Base* (KB), as well as RI are also defined. RI is any information required to render, process, interpret, use and understand data; for example, it may be a technical specification, or a data dictionary or a software tool. To preserve digitally encoded information over the long term the OAIS Model requires that information remain accessible, understandable and usable by a specified DC. A DC is a group of users or consumers for whom the data is being maintained. Within the OAIS Model, intelligibility of the data by the DC is of paramount importance and RI is a key concept in achieving this [13]. The Model identifies three main types of RI: structural, semantic and other.

In section 3, we describe the development of a registry/repository of RI (RRoRI) [18] which aims to make relevant RI available in a readily accessible manner to third parties. The work is heavily based on the ideas in the OAIS model; it centres on the notion that RI is critical to the long-term access of digital information [19], [20]. The current implementation of RRoRI is based on the use of standards (ebXML) and freely available registry/repository software (freebXML) with its associated JAXR interfaces. As explained in 3.2, access to RI by third parties is enabled through the use of two key concepts: *Curation Persistent Identifiers* (CPIDs) and descriptive *RI labels* [24].

The crystallography domain and the workflow of the NCS are then examined in order to identify significant RI. Procedures at the NCS indicate that a number of well-defined, sequential stages are readily identifiable. 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. We have found that the Crystallographic Information File (CIF) format is central to working with contemporary crystallography data as well as maintaining access to its information content in the future. CIF is used as a publishing format; as well as being structured and machine-readable, it is capable of describing the whole experiment and modelling process.

As a result, section 6 clarifies the relationship between various types of CIF RI, including structure (file format specification), semantic (data dictionaries) and other (software). These relationships are then used to develop an RI Network for the CIF format. Section 7, goes on to describe an ingest tool which allows RI to be input into RRoRI, as well

as describing the crystallography RI that has been submitted to RRoRI so far. A simple use case scenario, in section 8, describes how the RI stored in RRoRI may be used in order to gain access to the information content of a CIF instance by someone unfamiliar with that file format.

We conclude with a discussion of the role of RI in curating and maintaining access to crystallography data and pointers to further work:

- The range and quantity of RI required for even a simple collection of data is potentially enormous. It is therefore practical to develop a collaborative and shared approach to the problem. It would benefit the whole community if service providers and developers of work-up software (e.g. SHELXS, SHELXL, XPREP) were to provide and maintain comprehensive descriptions of their file formats; also the export of raw data in the draft standard imgCIF/CBF (Crystallographic Binary Format) [36], by crystallographic instrumentation software is recommended.
- Explicit recording of relevant RI in a central and managed registry/repository such as RRoRI ensures that the CIF file format can be understood well into the future by those working across different disciplines as well as providing intelligible long term access to crystallographers.
- In order to associate an RI Network with the CIF files stored in the eCrystals repository, it would be necessary to record a CPID in the metadata record for each CIF instance file. This CPID would act as a point of entry into RRoRI by pointing to an RI label.
- It is likely that RI in itself may not be sufficient to guarantee effective access and reuse of digital data in the future; additional metadata such as reference, provenance, context and fixity information will also need to be recorded and maintained.
- RI will itself need to be curated and maintained to provide trusted, authoritative and secure RI that allows users to rely on its authenticity and integrity; this could perhaps be overseen by the DCC.
- Long term curation of the contents of a registry/repository of RI would have to be guaranteed through adequate sustainability and succession planning, perhaps with an organisation of guaranteed longevity such as the NARA, The National Archives or The British Library.
- An alternative to relying on a generic, central registry/repository is for the crystallography discipline to develop its own RI registry/repository maintained by the community or a body such as the IUCr. Such a registry/repository would form part of a global and distributed network of RI. The web pages currently maintained by the IUCr, while certainly providing up-to-date information on the CIF file format, are at present suitable only for human access. A registry/repository modelled on the RRoRI would cater for automated machine processing.
- Furthermore, we can envisage that registries/repositories of RI would have a valuable role to play in the shared services infrastructure part of the JISC Information Environment, helping to provide convenient access to data for both research and learning.

Acknowledgements

The eCrystals Federation Project and the Digital Curation Centre (DCC) and are funded by the UK's Joint Information Systems Committee (JISC). The Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) Project is an Integrated Project co-financed by the European Union within the Sixth Framework Programme.

Thanks to Stephen Rankin and Brian MCIlwrath (DCC & CASPAR Projects, Science and Technology Facilities Council) for their help with the process of ingesting crystallography Representation Information into RRoRI.

Revision History

Date	Contributor	Contribution
14 th October 2008	M Patel	Set up structure and outline of report and associated work
16 th October 2008	M Patel	Added section on O AIS
4 th February 2009	M Patel	Added section on Representation Information
11 th February 2009	M Patel	Added section on RI and eCrystals
13 th February 2009	M Patel	Started adding sections on RRoRI
16 th February 2009	M Patel	Added section on CIF RI
17 th February 2009	M Patel	Some restructuring
18 th February 2009	M Patel	Work on Introduction and RI Labels
4 th March 2009	M Patel	Added section on NCS Workflow
7 th – 8 th April 2009	M Patel	Work on text of various sections
9 th April 2009	M Patel	Added screen shots of RRoRI
15 th April 2009	M Patel	Added executive summary + tidying up; circulated to Simon Coles and Liz Lyon for comment
20 th April 2009	M Patel	Corrected typos and general tidying up
5 th May 2009	M Patel	Incorporated comments from Liz Lyon
19 th May 2009	M Patel	Some tidying up

Contents

1. Introduction.....	7
2. The OAIS Reference Model	7
2.1 Important Concepts.....	8
2.2 Mandatory Responsibilities	9
2.3 Functional Model.....	9
2.4 Information Model.....	10
2.5. Representation Information	11
2.5.1 Structure Information.....	11
2.5.2 Semantic Information.....	12
2.5.3 Other Information	12
3. Registry/Repository of Representation Information.....	12
3.1 RRoRI Implementation.....	13
3.2 Curation Persistent Identifiers and RI Labels	13
4. The Crystallography Domain.....	13
5. EPSRC UK National Crystallography Service.....	14
5.1 Crystal Structure Determination Workflow.....	15
6. Crystallography Representation Information.....	16
7. Populating RRoRI.....	17
8. RI Network Usage Scenario.....	19
9. Discussion & Further Work.....	20
References.....	21

1. Introduction

The eCrystals Federation project is concerned with setting up a federation of institutional repositories for the management and dissemination of the raw, derived and results data from crystallographic experiments [1]. It builds on the work of the eBank-UK project [2] which developed and implemented the eCrystals repository [3], focusing on the workflows of the laboratory based experimental technique of chemical crystallography undertaken at the EPSRC National Crystallography Centre (NCS) based in Southampton. Following the creation of a completed crystal structure determination, data is uploaded into eCrystals and supplemented with chemical and bibliographic metadata. A subsequent scoping study, undertaken as part of phase 3 of the eBank-UK project identified several issues pertinent to the curation and preservation of crystallography data [4], amongst them was the importance of the concepts underlying the OAIS Reference Model [5] and its associated notion of *Representation Information* (RI).

Consequently, this report is concerned with an investigation of RI for crystallography data and its role in the curation, maintenance and management of such data. We begin with a brief overview of aspects of the OAIS Reference Model that are relevant to this study. This is followed by a description of the implementation of a registry/repository for storing RI. The crystallography domain and the workflow of the NCS are then examined in order to identify significant RI and its ingest into the registry/repository.

2. The OAIS Reference Model

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.

The Reference Model establishes a conceptual framework of terms and components for use in the preservation of information and was never intended to prescribe implementation. It identifies the environment within which an OAIS operates as well as its basic functions, defining: a functional model; an information model and an information flow model. Within the preservation community, it has established itself as an important standard, influencing: the development of preservation metadata [6]; architectures and systems design of repositories [7]; as well as conformance criteria for archival repositories [8]. The Model recommends the setting up of conformance and certification processes and has been used as a foundation and starting point by several groups working in this area (notably the RLG-NARA [9] and NESTOR work on trusted digital repositories [10]). The notion of the trustworthiness of an archive or repository in the context of the eCrystals Federation project is further discussed in a sister report [11].

One of the characteristics of the OAIS Reference Model is that it highlights the fact that curation and preservation are continuous processes which in the extreme may require infinite attention, hence the Model's emphasis on continual monitoring of the environment within which the OAIS operates.

2.1 Important Concepts

The Reference Model considers long term preservation to be the act of maintaining information, in a correct and independently understandable form, over the long term. Long term is defined to be long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Several concepts form the basis for understanding the workings of an OAIS:

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

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

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

Information Object: results from *Representation Information* being applied to a data object. It is important to appreciate that the OAIS model is concerned with preserving both the meaning and reusability of an information object.

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

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

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

Preservation Description Information (PDI): is any metadata deemed of particular relevance to the curation and preservation of the content information in an OAIS. The standard mentions four specific types:

- *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).

Matters relating to the development of preservation metadata within the context of the eCrystals Federation project are further discussed in a sister report [12].

2.2 Mandatory Responsibilities

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

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure its long-term preservation
- Determining the DC
- Ensuring that information provided by the OAIS is understandable by the DC without having to refer back to the producer of the information
- Following documented policies and procedures
- Making the preserved information available to the DC

However, our main concern in this study relates to the concept of *Representation Information* and the OAIS functional and information models.

2.3 Functional Model

The organisational and operational environment of the OAIS model is set in the context of producers (who generate the information to be archived), consumers (who retrieve the information) and management (the wider organisation hosting the OAIS). The main components of an OAIS are modelled as six functional entities, as shown in Figure 1.

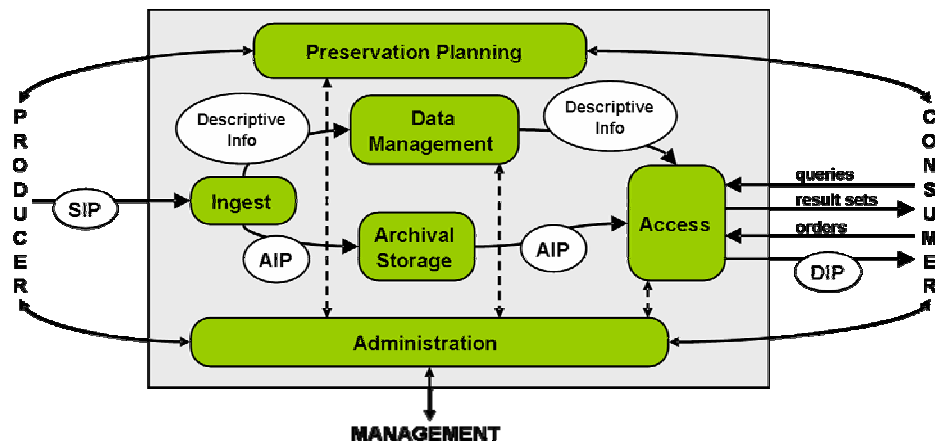


Figure 1: OAIS Functional Entities reproduced from Figure 4-1 in the OAIS Reference Model [5]

- *Ingest*: services and functions that accept SIPs from Producers; prepare AIPs for storage and ensure that AIPs and their supporting Descriptive Information become established within the OAIS
- *Archival Storage*: services and functions used for the storage and retrieval of AIPs
- *Data Management*: services and functions for populating, maintaining, and accessing a wide variety of information
- *Administration*: services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis

- *Preservation Planning*: services and functions for monitoring the OAIS environment and ensuring that content remain accessible to the DC.
- *Access*: services and functions which make the archival information holdings and related services visible to Consumers (e.g. search and retrieval)

Information objects move through an OAIS in an encapsulated form known as an information package which includes the Data Object and its RI, together with PDI (see section 2.1). Information packages come in three varieties; archival; submission and dissemination. The archival information package (AIP) is the version that the OAIS actually preserves. Information is submitted to the OAIS in submission information packages (SIPs) and dissemination information packages (DIPs) are versions of AIPs tailored to consumer requirements.

2.4 Information Model

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

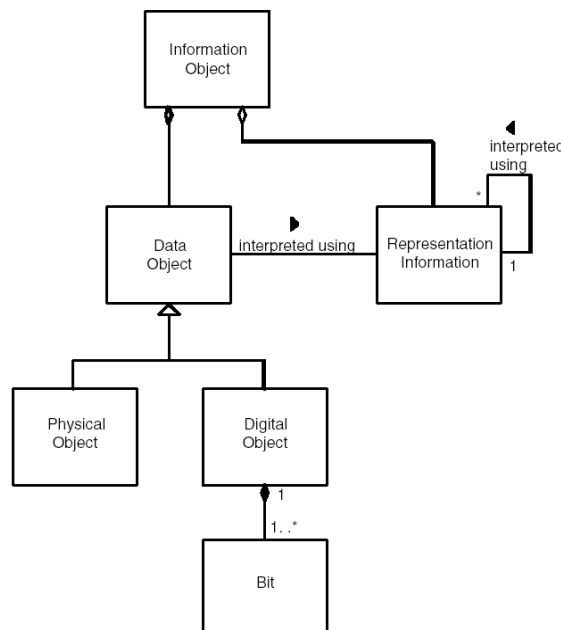


Figure 2: An OAIS Information Object, reproduced from Figure 4-10 in the OAIS Reference Model [5]

To preserve digitally encoded information over the long term the OAIS Model requires that information remain accessible, understandable and usable by a specified DC. A DC is a group of users or consumers for whom the data is being maintained. Within the OAIS Reference Model, intelligibility of the data by the DC is of paramount importance and RI is a key concept in achieving this [13].

2.5. Representation Information

RI is defined as whatever is needed to allow a Data Object to be converted to an Information Object. As explained earlier, RI can comprise any information that is required to render, process, interpret, use and understand digital data, including: file formats, software, algorithms, standards, semantic information etc. RI is recursive in nature enabling Representation Information Networks to be built up; using one element of RI in a meaningful manner may well require further RI. It is expected that the recursion will terminate for a particular DC when the RI can be understood using that designated community's KB. A problem with RI is that the amount needed for a particular object could be vast and impractical to collect in reality. It is for this reason that the concept of the designated community is so important; it enables a limit to be placed on the amount of RI which it is necessary to capture at any particular time. It is essential that RI itself is curated and preserved to maintain access to other digital data.

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

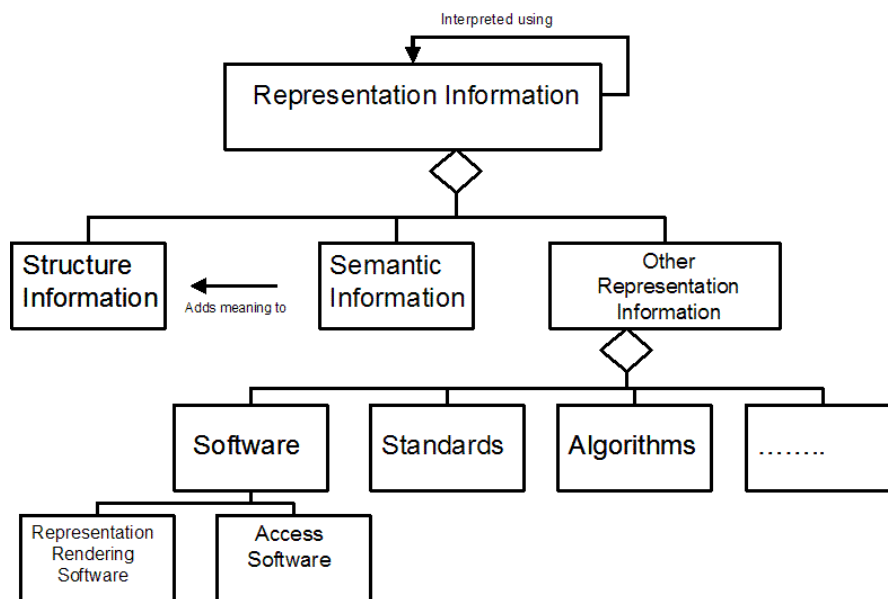


Figure 3: Types of Representation Information [5]

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

2.5.1 Structure Information

In a digital world, structural information manifests itself largely in the form of digital file formats for text, images, audio, moving images, datasets, 3D models as well as time-varying or dynamic data. It is useful to distinguish between formats which are used mainly for rendering (for human consumption) and formats used for automated processing [7]. The

former include many commercially based formats such as the succession of Microsoft Word formats; the details of such formats are likely to be proprietary and difficult or impossible to obtain; in this case, the original software, or some equivalent application, may be required to enable access. The latter are more likely to be simpler, with open source access software. Community standards in the use of structural information play an important part in establishing the knowledge base of a particular designated community.

Formal descriptions of file formats are useful in enabling automated processing, for example using the EAST [14], FLAVOR [15], or DFDL [16] languages.

2.5.2 Semantic Information

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

2.5.3 Other Information

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

3. Registry/Repository of Representation Information

The Digital Curation Centre (DCC) [17] and the CASPAR Project [7] are developing a registry/repository of RI (RRoRI) [18]. This is intended to be an authoritative source of RI for the community responsible for the collection, curation and management of data. The primary function is to provide and share information that enables managers of digital information to make informed decisions with regards to curation strategies. The RRoRI aims to make relevant RI available in a readily accessible manner to third parties. The work is heavily based on the ideas in the OAIS model; it centres on the notion that RI is critical to the long-term access of digital information [19], [20]. Collection and maintenance of suitable RI mitigates the difficulties related to the preservation of understandable information - data formats, software, standards and programming languages become obsolete; the documentation for these is often poor or non-existent; and the specialised knowledge needed to manipulate the data often disappears with time.

The huge burden of collecting and maintaining adequate RI requires that a global, distributed network of RI be developed; to this end the proponents of RRoRI are currently collaborating with projects such as PRONOM [21] (developed by the National Archives) and the Global Digital Format Registry (GDFR), developed by the Digital Library Federation (DLF). PRONOM and the GDFR have now joined forces to form a new initiative known as the Unified Digital Formats Registry (UDFR) [22] which will complement the more wide

ranging RI in RRoRI, by focusing on the provision of details about file formats (essentially the Structure type of RI) [23].

3.1 RRoRI Implementation

The current implementation of RRoRI [18] is based on the use of standards (ebXML) and freely available registry/repository software (freebXML) with its associated JAXR interfaces. In addition, the provision of an abstraction layer in terms of an API provides independence from the JAXR/ebXML-specific implementation.

To support use of pre-existing RI, RRoRI is able to handle multiple classification schemes as well as that of the OAIS Reference Model. In addition, the OAIS classification of RI has been expanded to cater for finer granularity in categorising different types of RI. For example, at present, the OAIS category of semantic RI has been subdivided into: Data, Document, Language, Models and Standards; other RI has been subdivided into: AccessSoftware, Algorithms, CommonFileTypes, ComputerHardware, ProcessingSoftware, RepresentationRenderingSoftware, Media, Physical and Software. Several of these categories themselves have further subdivisions e.g. Data has a DictionarySpecification as a sub-type.

3.2 Curation Persistent Identifiers and RI Labels

Access to RI by third parties is enabled through the use of two key concepts: *Curation Persistent Identifiers (CPIDs)* and descriptive *RI labels* [24] (see Figure 6 for an example). A CPID (currently implemented as a UUID) is a unique identifier for an information object which may be an item of RI in the registry. An RI label, comprising an XML schema, provides a mechanism for describing and structuring multiple elements of RI which relate to a particular digital object, with each having its own CPID as an entry point into RRoRI. In this manner, a digital object can be associated with a label, which points to items of RI. Those items of RI may in turn point to further items of RI to create a recursive structure. The recursion is terminated when the item of RI refers to an assumption about a defined KB. The RI Network for the original object therefore comprises the set of RI items resulting from recursively following the pointers from its label. Ideally, the operation of retrieving the RI relevant to a digital object will be automated and transparent to the end user.

4. The Crystallography Domain

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. This 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 data is in the order of Megabytes and results data is normally Kilobytes in terms of size.

In terms of current practice, the crystallography community takes a relatively organized approach to the management of their 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. The

Crystallography Information File (CIF) is the de facto exchange standard [25]. 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; its Crystal Structure Database (CSD) 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 [26] 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 [27], developed at the Unilever Centre for Molecular Informatics (University of Cambridge), provide open access to aggregated CIF data through the use of web-crawlers.

5. EPSRC UK National Crystallography Service

As part of the eBank-UK project, the EPSRC UK National Crystallography Service (NCS) has constructed an institutional data repository (eCrystals [3]), to provide open access and rapid dissemination of derived and results data from crystallography experiments, as well as linking research data to publications and scholarly communication [2].

The figure displays two screenshots of a Data Structure Report from the eCrystals website. The left screenshot shows the main report page for the compound 2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclophosphazene. It includes a navigation menu, sample originator information (D.B. Davies, R.A. Shaw, A. Kikic, M. Odyta, and A. Uski), data collection details (S.J. Coles, L.S. Hubb, and M.E. Ligr), structure determination details (S.J. Coles, J.S. Rutherford, and M.B. Hursthouse), and a ball-and-stick model of the molecule. The right screenshot shows a detailed list of available files, categorized by type, with file names and sizes. The categories include Data collection parameters, Refinement, Solution, Processing, Refinement results, and Other Files.

File Name	Size
2005qc0007_checkcif.htm	9k
2005qc0007_res	5k
2005qc0007_xl.lst	29k
2005qc0007_prp	5k
2005qc0007_xs.lst	44k
2005qc0007_hk	532k
2005qc0007_hm	11k
2005qc0007_0k.jpg	91k
2005qc0007_h0.jpg	87k
2005qc0007_hk0.jpg	79k
2005qc0007_crystal.jpg	17k
2005qc0007.doc	186k
2005qc0007.lct	138k

Figure 4: An example Data Structure Report in eCrystals [28].

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

eCrystals website [3].

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

5.1 Crystal Structure Determination Workflow

Crystallography is concerned with determining the structure of a molecule and its three dimensional orientation with respect to other molecules in a crystal by analysis of diffraction patterns obtained from X-ray scattering experiments. In each experiment, the process relates to the determination of one structure, comprising both the molecular connectivity and the packing arrangements between molecules in the crystal being examined. The final result is a crystal structure in the form of a CIF file.

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 5. 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.

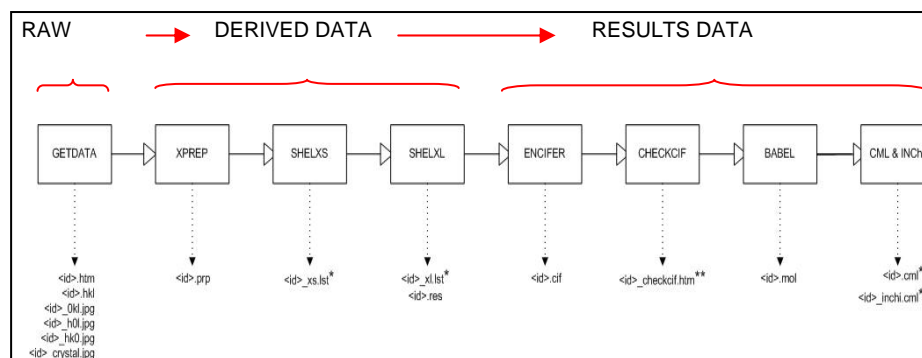


Figure 5: Workflow model of the EPSRC UK National Crystallographic Centre (NCS)

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 4 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 [25] 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. The CIF itself cannot provide a reference back to the raw data file {.hkl}, but this data can be presented in a text dictionary driven form i.e. with a CIF header {.fcf}. 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 [29].

Another type of file format included in the Final Result is a Chemical Markup Language (CML) encoding [30]. The CML file is translated from the CIF and introduces complementary 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 [31], 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 [32] software obtainable from SourceForge.

The data collection, processing and solution stages are the main areas involving the work-up of the original data. The data collection stage provides JPEG files as representations of the raw data, but also proprietary formats generated by specific instrumentation 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 [33]. 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 4, a 3D graphical rendition of the molecule is also generated using `jmol`; this can be rotated interactively on the eCrystals website.

6. Crystallography Representation Information

We have initially chosen to examine the RI network associated with the CIF file format, since this appears to be a critical format in the designated community at present. The CIF file format is central to working with contemporary crystallography data as well as maintaining access to its information content in the future.

For data stored in a CIF file, understanding the format is an essential but preliminary step towards interpretation of the underlying information object, since the CIF format is essentially a container, interpretation of the content requires additional RI, such as the CIF core data dictionary. In addition, there are numerous software utilities currently in use for checking the syntactical validity of a CIF file e.g. CheckCIF and `vcif`. The IUCr supports developments relating to the CIF file format and maintains a web page which provides up-to-date information [25]. All of these types of information can be considered to be critical RI for the interpretation of CIF data files and should be collected together into an RI Network. Figure 6 shows part of the RI Network for the CIF file format, including examples of structure, semantic and other information.

Space limitations and the recursive nature of RI networks mean that we are unable to reproduce the entire RI Network here. However, a more complete (textual) version is available on the Web [34] providing an indication of the complexity and granularity of the information required. The RI Network in Figure 6 shows that a CPID pointing to an RI label is associated with a specific CIF data file.

The RI label is stored within RRoRI and contains CPIDs pointing to structure, semantic and other RI. The structural RI shown is that of the CIF file format specification and a dictionary definition language for the format; semantic information is provided by a CIF core data dictionary, which can be supplemented with further sub-domain specific extensions, such as the *powder*, *rho* and *symmetry* dictionaries; and other RI is included in the

form of two software tools, a CIF syntax checker and a conversion utility. Each of these pieces of RI is a digital object in its own right and points to a further RI label which describes its own associated RI Network.

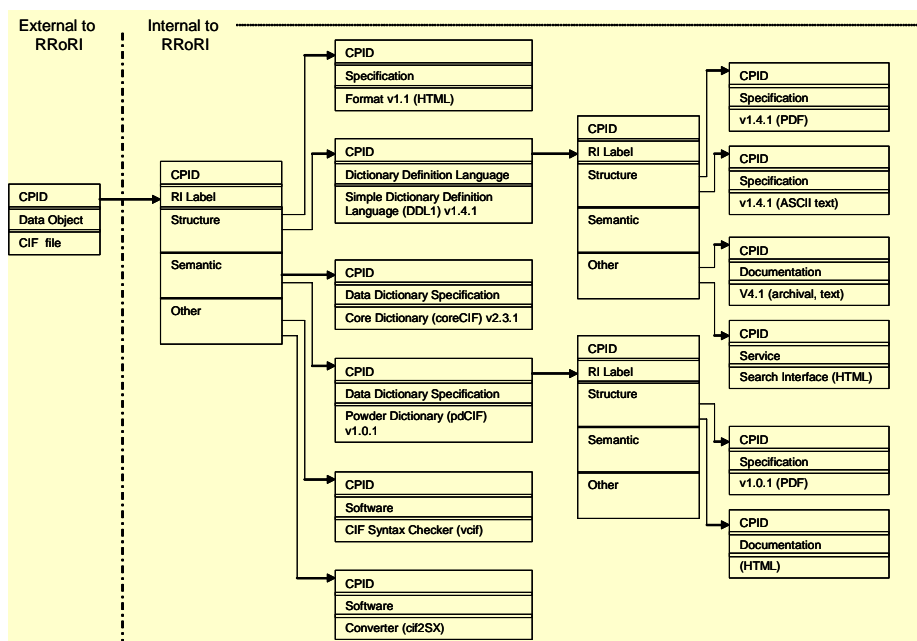


Figure 6: Graphical visualisation of part of an RI Network for the CIF file format [34].

7. Populating RRoRI

An ingest tool with a GUI serves as a client to the RRoRI server and enables RI to be input into RRoRI (Figure 7). This utility also facilitates the maintenance of RI and is particularly useful as a search interface allowing RI already in the registry to be identified and reused helping to avoid duplication, share resources, coordinate access and minimise effort.

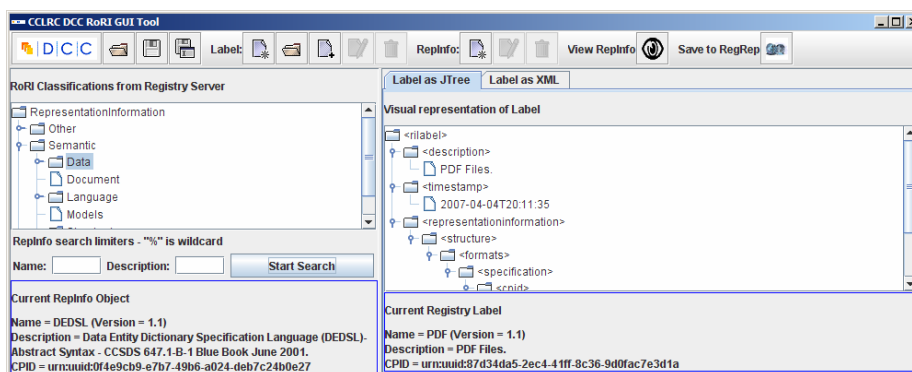


Figure 7: A client with a GUI for ingest, search and retrieval of RI and RI labels.

Figure 8 shows a typical dialogue which allows third parties (access rights permitting) to populate RRoRI with RI. Such a tool and interface is necessary because domain expertise is required in order to identify and record suitable and adequate RI.

Enter new Representation Information.....

File or URL:

Classification:

RepInfo Label:

Name:

Description:

Mime type:

Figure 8: A dialogue for the process of RI ingest.

Representation Information Registry Repository Home Frequently Asked Questions Documentation User Guide

Current User: Registry Guest

Tasks Search Explore Versioning ON

Content Language:

Registry Objects Registry Objects Help

Deletion Option:

Status Type

Results 26 - 43 of 43 Bookmark And Relate Help

Pick	Details	Object Type	Name	Description	Version	Version Comment	Status
<input type="checkbox"/>	Details	Document	CIF Powder Dictionary	Documentation CIF Powder Dictionary	1.1		Submitted
<input type="checkbox"/>	Details	Document	CIF Symmetry Dictionary	Documentation CIF Symmetry Dictionary	1.1		Submitted
<input type="checkbox"/>	Details	Document	CIF Core Data Dictionary	Documentation for CIF Core Data Dictionary	1.1		Submitted
<input type="checkbox"/>	Details	Document	Image CIF Dictionary	Documentation Image CIF Dictionary	1.1		Submitted
<input type="checkbox"/>	Details	Document	CIF Macromolecular Dictionary	Documentation CIF Macromolecular Dictionary	1.1		Submitted
<input type="checkbox"/>	Details	AccessSoftware	enCIFer_1.3.exe	enCIFer for Windows	1.1		Submitted
<input type="checkbox"/>	Details	AccessSoftware	enCIFer_1.3.dmg	enCIFer for Macintosh	1.1		Submitted
<input type="checkbox"/>	Details	AccessSoftware	encifer_1.3.1_linux.tar.gz	enCIFer for Linux	1.1		Submitted
<input type="checkbox"/>	Details	Document	encifer_user_guide	EnCIFer GUI editor for single or multi-block CIFs	1.1		Submitted
<input type="checkbox"/>	Details	Dictionary	mmcif_ddl_2.1.6	DDL2 - relational Dictionary Definition Language for CIF	1.1		Submitted
<input type="checkbox"/>	Details	AccessSoftware	vcif.exe	CIF syntax checker - MSDOS version	1.1		Submitted
<input type="checkbox"/>	Details	AccessSoftware	vcif.tar.Z	CIF syntax checker - Linux version	1.1		Submitted
<input type="checkbox"/>	Details	Document	vcif Documentation	Syntax checker for CIF	1.1		Submitted
<input type="checkbox"/>	Details	RepInfoLabel	Label for cif_sym	Symmetry dictionary for CIF	1.1		Submitted
<input type="checkbox"/>	Details	Dictionary	cif_sym	Symmetry dictionary for CIF	1.2		Submitted
<input type="checkbox"/>	Details	RepInfoLabel	Label for (IUCr) CIF 1.1 specification-syntax	(IUCr) CIF 1.1 specification-syntax - zipped web archive	1.1		Submitted
<input type="checkbox"/>	Details	DictionarySpecification	(IUCr) CIF 1.1 specification-syntax	(IUCr) CIF 1.1 specification-syntax - zipped web archive	1.2		Submitted
<input type="checkbox"/>	Details	RepInfoLabel	Label for (IUCr) CIF 1.1 specification-syntax	(IUCr) CIF 1.1 specification-syntax - zipped web archive	1.2		Submitted

Previous **1** 2 Next

About Representation Information
Copyright 2001-2006, DCC, CASPAR and FreebXML
About Registry

Figure 9: Part results of a search on RRoRI for all RI relating to the CIF file format (Page 2).

The screen capture in Figure 9 shows part of the results of a search in RRoRI for all the RI relevant to the CIF file format. At present this includes the following:

RepInfoLabel	Label for (IUcr) CIF 1.1 specification-syntax	(IUcr) CIF 1.1 specification-syntax
DictionarySpecification	(IUcr) CIF 1.1 specification-syntax	(IUcr) CIF 1.1 specification-syntax - zipped web archive
Dictionary	mmcif_ddl_2.1.6	DDL2 dictionary for CIF files Version 2.1.6
Dictionary	cif_core-2.3.1	CIF file format core dictionary Version 2.3.1
Dictionary	cif_img	Image Dictionary for CIF
Dictionary	cif_mm	Macromolecular dictionary for CIF Version 2.0.09
Dictionary	cif_ms	Modulated and Composite structures for CIF V1.0.1
Dictionary	cif_pd	Powder dictionary for CIF V1.0.1
Dictionary	cif_rho	Electron density dictionary for CIF V1.0.1
Dictionary	cif_sym	Symmetry dictionary for CIF
Document	CIF Core Modulated Dictionary	Documentation CIF Core Modulated Dictionary
Document	CIF Electron Density Dictionary	Documentation CIF Electron Density Dictionary
Document	CIF Powder Dictionary	Documentation CIF Powder Dictionary
Document	CIF Symmetry Dictionary	Documentation CIF Symmetry Dictionary
Document	CIF Core Data Dictionary	Documentation for CIF Core Data Dictionary
Document	Image CIF Dictionary	Documentation Image CIF Dictionary
Document	CIF Macromolecular Dictionary	Documentation CIF Macromolecular Dictionary
AccessSoftware	enCIFer_1.3.exe	enCIFer for Windows
AccessSoftware	enCIFer_1.3.dmg	enCIFer for Macintosh
AccessSoftware	encifer_1.3.1_linux.tar.gz	enCIFer for Linux
Document	encifer_user_guide	EnCIFer GUI editor for single or multi-block CIFs
Dictionary	mmcif_ddl_2.1.6	DDL2 - relational Dictionary Definition Language for CIF
AccessSoftware	vcif.exe	CIF syntax checker - MSDOS version
AccessSoftware	vcif.tar.Z	CIF syntax checker - Linux version
Document	vcif Documentation	Syntax checker for CIF
RepInfoLabel	Label for cif_sym	Symmetry dictionary for CIF
Dictionary	cif_sym	Symmetry dictionary for CIF
RepInfoLabel	Label for (IUcr) CIF 1.1 specification-syntax	(IUcr) CIF 1.1 specification-syntax - zipped web archive
DictionarySpecification	(IUcr) CIF 1.1 specification-syntax	(IUcr) CIF 1.1 specification-syntax - zipped web archive
RepInfoLabel	Label for (IUcr) CIF 1.1 specification-syntax	(IUcr) CIF 1.1 specification-syntax - zipped web archive

8. RI Network Usage Scenario

Much scientific data is now “born-digital” and relies heavily on software applications for processing, access and rendering. The complexity and granularity of the information accumulated within an RI Network makes it important to provide an automated traversal of the network; this is supported through the use of the CPID.

We can envisage a scenario in which a user downloads a CIF file consisting of a crystal structure from an archive (e.g. the eCrystals repository); the metadata record of the data file contains a CPID. If the user is unfamiliar with the file format and the dataset s/he can use the CPID to request RI from RRoRI (or a distributed global network of RI). The set of RI that the user receives would include: the file format specification for the CIF file format; the CIF core data dictionary; and perhaps the CheckCIF software utility. Each of these pieces of RI would be a digital object itself which may well have its own CPID in case the user requires further RI in order to understand and reuse the CIF data file.

9. Discussion & Further Work

A comprehensive discussion of issues relating to the use of RI in digital curation is provided in [35]. Here we highlight several areas of particular significance in managing and maintaining access to crystallography data.

As we have seen, crystal structure determination typically involves a pipeline of digital processes (see section 5.1). Consequently, the range and quantity of RI required for even a simple collection of data is potentially enormous. It is therefore practical to develop a collaborative and shared approach to the problem. It would benefit the whole community if service providers and developers of work-up software (e.g. SHELXS, SHELXL, XPREP) were to provide and maintain comprehensive descriptions of their file formats; also the export of raw data in the draft standard imgCIF/CBF (Crystallographic Binary Format) [36], by crystallographic instrumentation software is recommended.

Explicit recording of relevant RI in a central and managed registry/repository such as RRoRI ensures that the CIF file format can be understood well into the future by those working across different disciplines as well as providing intelligible long term access to crystallographers.

In order to associate an RI Network with the CIF files stored in the eCrystals repository, it would be necessary to record a CPID in the metadata record for each CIF instance file. This CPID would act as a point of entry into RRoRI by pointing to an RI label stored within the registry/repository.

It is likely that RI in itself may not be sufficient to guarantee effective access and reuse of digital data in the future; additional metadata such as the Preservation Description Information (PDI) of the OAIS Reference Model will be needed to provide supplementary information. PDI is any metadata deemed of particular relevance to the curation and preservation of the content information in an OAIS and includes reference, provenance, context, and fixity information, as discussed in section 2.1.

Given that the information contained in RRoRI is vital for long term access to crystallography CIF files, the associated RI will itself need to be curated and maintained to provide trusted, authoritative and secure RI that allows users to rely on its authenticity and integrity, perhaps overseen by the DCC. In addition, long term curation of the contents of RRoRI would have to be guaranteed through adequate sustainability and succession planning, perhaps with an organisation of guaranteed longevity such as the NARA, The National Archives or The British Library.

An alternative to relying on a generic, central registry/repository is for the crystallography domain to develop its own RI registry/repository maintained by the community or a body such as the IUCr. Such a registry/repository would form part of a global and distributed network of RI. The web pages currently maintained by the IUCr, while certainly providing up-to-date information, are at present suitable only for human access. A registry/repository modelled on the RRoRI would cater for automated machine processing.

Furthermore, we can envisage that registries/repositories of RI would have a valuable role to play in the shared services infrastructure part of the JISC Information Environment [37], helping to provide convenient access to data for both research and learning.

References

1. eCrystals Federation Project, http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page
2. Duke, M., Day, M., Heery, R., Carr, L., Coles, S.: Enhancing access to research data: the challenge of crystallography, Proceedings JCDL'05, Denver, Colorado, USA, 2005
3. The Crystal Structure Report Archive –eCrystals Data Repository, <http://ecrystals.chem.soton.ac.uk>
4. 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
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%20mode%201%22>
6. PREMIS Data Dictionary for Preservation Metadata version 2.0, Preservation Metadata Maintenance Activity, 2008, <http://www.loc.gov/standards/premis/>
7. Giaretta, D.: The CASPAR Approach to Digital Preservation, International Journal of Digital Curation, Vol. 2 (1) 2007
8. Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Version 1.0, Center for Research Libraries and RLG Programs, 2007
9. *Trusted Digital Repositories: Attributes and Responsibilities*, An RLG OCLC Report, May 2002, <http://www.rlg.org/legacy/longterm/repositories.pdf>
10. Dobratz, S. and Schoger, A.: *Digital Repository Certification: A Report from Germany*, DINI/NESTOR, October 2005
<http://edoc.huberlin.de/oa/articles/reh7CbxRopdUA/PDF/23yn183UoMBU.pdf>
11. Patel M.: Preservation Planning for Crystallography Data, eCrystals Federation Project, WP4, April 2009 (in progress)
12. Patel M.: [Preservation Metadata for Crystallography Data](#), eCrystals Federation Project, WP4, April 2009 (in progress)
13. Tzitzikas, Y.: On Preserving the Intelligibility of Digital Objects through Dependency Management, Proceedings International Conference PV'2007, Oberpfaffenhofen/Munich, Germany, 2007
14. The EAST Data Description Language, <http://east.cnes.fr/english/index.html>
15. Eleftheriadis, A., Hong, D.: Flavor: a formal language for audio-visual object representation, Proceedings 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004
16. Data Format Description Language, <http://forge.gridforum.org/projects/dfdl-wg/>
17. The Digital Curation Centre (DCC), <http://www.dcc.ac.uk/>
18. Registry/Repository of Representation Information (RRoRI), <http://registry.dcc.ac.uk/>
19. Giaretta D., Patel M., Rusbridge A., Rankin S., McIlwrath B.: Supporting e-Research Using Representation Information, Proceedings UK e-Science All Hands Meeting, 2005, <http://www.allhands.org.uk/2005/proceedings/papers/447.pdf>
20. Giaretta D., Rankin S., McIlwrath B., Rusbridge A., Patel M.: Representation Information for Interoperability Now and with the Future, Proceedings MSST 2005, pp54-58, International IEEE Symposium on Mass Storage and Systems, 20-24th June

- 2005, Sardinia, Italy, <http://www.soe.ucsc.edu/~elm/msst05/MSST05-Sardinia-Proceedings-2.pdf>
21. PRONOM Registry of Technical Information, The National Archives, UK, <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>
 22. Unified Digital Formats Registry, <http://www.gdfr.info/>
 23. Brown, A.: PLANETS White Paper, Representation Information Registries, 2008 http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf
 24. DCC Development Team: DCC Label Report, <http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCInfoLabelReport>
 25. CIF -The Crystallographic Information File, <http://www.iucr.org/iucr-top/cif/>
 26. The ReciprocalNet Consortium, <http://www.reciprocalnet.org/>
 27. CrystalEye, Unilever Centre for Molecular Informatics, University of Cambridge, UK <http://wwmm.ch.cam.ac.uk/crystaleye/index.html>
 28. Example eCrystals Data Structure Archive Report, <http://ecrystals.chem.soton.ac.uk/300/>
 29. IUCr checkCIF validation service, <http://checkcif.iucr.org/>
 30. Chemical Markup Language (CML), <http://www.ch.ic.ac.uk/rzepa/cml/>
 31. International Chemical Identifier (INChI), <http://www.inchi.info/>
 32. Open Babel: The Open Source Chemistry Toolbox, http://openbabel.sourceforge.net/wiki/Main_Page
 33. SHELXS software suite, <http://shelx.uni-ac.gwdg.de/SHELX/>
 34. Example Representation Information Network for CIF file format, <http://homes.ukoln.ac.uk/~lismp/ECDL2009/RINetCIF.html>
 35. Patel, M., Ball, A: Challenges and Issues Relating to the Use of Representation Information for the Digital Curation of Crystallography and Engineering Data, The International Journal of Digital Curation, Vol.3 (1) 2008 <http://www.ijdc.net/index.php/ijdc/article/viewFile/64/43>
 36. The Crystallographic Binary File Format -DRAFT PROPOSAL, http://www.esrf.eu/computing/Forum/imgCIF/cbf_definition.html
 37. JISC Information Environment, <http://www.jisc.ac.uk/whatwedo/themes/informationenvironment.aspx>