The Role of OAIS Representation Information in the Digital Curation of Crystallography Data

Manjula Patel[1], Simon Coles[2], David Giaretta[3], Stephen Rankin[3], Brian McIlwrath[3]

[1]UKOLN, University of Bath, UK
[2] EPSRC NCS, University of Southampton, UK
[3] Science & Technology Facilities Council, UK
{m.patel@ukoln.ac.uk, s.j.coles@soton.ac.uk, david.giaretta@stfc.ac.uk,
stephen.rankin@stfc.ac.uk, brian.mcilwrath@stfc.ac.uk}

Abstract

Reusable high quality data are emerging as the raw material of contemporary e-science. Large volumes of scientific data are now "born-digital" and need to be curated to facilitate use and reuse. Representation Information (RI) as defined by the OAIS Reference Model is increasingly recognised as being vital to the long term curation and preservation of meaningful and reliable digital data. This paper is concerned with an investigation of RI for crystallography data and its role in the curation, maintenance and management of such data. We describe how the explicit recording of relevant RI can facilitate long term access and maintain intelligibility of the Crystallographic Information File format (a critical file format in the crystallography domain).

1. Introduction

Contemporary e-science is one of the most significant producers of large volumes of digital data. Scientists are now able to collect huge amounts of data due to advances in instrumentation, data storage technologies, computational power, improvements in algorithms and the development of grid and cyber infrastructures [1], [2]. The result is that the actual nature of science is changing and becoming increasingly reliant on the mining and analysis of large databases of data such as the Protein Data Bank and GenBank. High quality data suitable for repurposing are therefore emerging as the raw material of e-science. They are used in innovative scientific endeavor (such as predictive science) as well as in the verification, validation and replication of scientific results. According to Carlson, "To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data" [3].

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], [5]. 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, organizational, financial and legal as well as technical issues.

Nonetheless, technological obsolescence of hardware, software and file formats, poses a major threat to digital information; data can become inaccessible within a very short period of time. Moreover, much digital information is dependent on specific configurations of hardware and software applications in addition to details of the semantics associated with the data in order to make them understandable and usable by humans.

A variety of techniques have been proposed and explored to combat the effects of rapidly changing technologies and media degradation: bit-stream copying, refreshing, the use of durable media, digital archaeology and replication. Strategies aimed at preserving access to the information content and providing functional preservation include: technology preservation, analogue backups, migration, normalization, emulation and encapsulation. One particular strategy concerned with mitigating the effects of technology evolution is based on the use of Representation Information (RI); the *Reference Model for an Open Archival Information System (OAIS)* [6] identifies the critical role that RI plays in maintaining the accessibility, understandability and usability of the information content of digital data. RI is basically any

978-0-7695-3877-8/09 $26.00 © 2009 Crown Copyright
DOI 10.1109/e-Science.2009.27

IEEE computer society

information that is required to render, interpret, process and understand data; it includes, file formats, software, algorithms, standards and semantic information (including contextual information relating to the process of data collection). It is increasingly being recognised as essential for both curation and contemporary use of digital data.

This paper is concerned with an investigation of RI for crystallography data and the role it plays in the curation, maintenance and management of such data. Crystallography is the sub-discipline of chemistry concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal through the analysis of diffraction patterns obtained from X-ray scattering experiments. 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 are in the order of Gigabytes, derived data in the order of Megabytes and results data are normally Kilobytes in size. In terms of data formats, they range from proprietary (binary) through to highly structured dictionary defined text.

We begin by examining the OAIS Reference Model and its concept of RI, followed by a description of the development of a Registry/Repository for RI (RRoRI) which aims to provide an infrastructure for the management of RI. We then provide an analysis of the workflow used at the EPSRC UK National Crystallographic Service in order to identify the most significant RI for contemporary crystallography data. The structuring of RI related to the Crystallography Information File (CIF); it's ingest into RRoRI and an example usage scenario precedes a discussion and concluding comments.

2. OAIS and Representation Information

Development of the OAIS Reference Model has been led by the Consultative Committee for Space Data Systems (CCSDS). It was adopted as an ISO standard in 2003 (ISO 14721:2003). 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 OAIS, so the model is equally applicable to dark as well as open archives. The model has recently undergone an open review process and a revision is imminent.

The OAIS Model establishes a conceptual framework of terms and components for use in the preservation of information, but does not prescribe implementation. It identifies the environment within which an OAIS operates as well as its basic functions within the context of producers, consumers and the management of data. The Model has achieved widespread adoption, influencing the development of preservation planning [7], preservation metadata [8], architectures and systems of repositories [9] as well as conformance and certification criteria for archives [10]. One of the characteristics of the Model is that it highlights the fact that curation and preservation are continuous processes which may require attention into the indefinite future; hence the Model's emphasis on continual monitoring of the environment within which the OAIS operates. To preserve digitally encoded information over the long term the model requires that information remain accessible, understandable and usable by a specified *Designated Community (DC)*. A DC is a group of users or consumers for whom the data are 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 [11].

2.1 Representation Information

Digital data inherently require additional information and methods to convert them into a form that can be interpreted for use and reuse. Fundamentally, RI is everything that is needed to make a particular collection of bits understandable and usable. Information in the OAIS Model is regarded as being a combination of Data and RI; the process of applying RI to a *Data Object* (bit-stream) yields an *Information Object* which allows for the full interpretation of the data into meaningful information.

Since comprehensive RI is needed to preserve access to information, we need to understand the variety of forms RI may take. The Reference Model considers three basic types of RI: *structure, semantic* and *other*.

Structure information manifests itself largely in the form of digital file formats for text, images, audio, moving images, datasets and 3D models as well as time-varying or dynamic data. It is useful to distinguish between formats which are used predominantly for rendering (i.e. for human consumption) and formats that are used for automated processing. 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 facilitate access. Structure information could be simple text but formal descriptions of file formats are useful in enabling automated processing, for example through the use of languages such as EAST [12], FLAVOR [13] or DFDL [14].

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

Other 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).

2.2 Representation Information Networks

In many cases, any particular piece of RI may also be a digital object which itself needs its own RI, thus creating a Representation Information Network; such networks are recursive in nature and in the extreme may well require an infeasible amount of RI to be recorded. For practicality, the notion of the DC and its associated *Knowledge Base (KB)* can be used to place a limit on the amount of RI that needs to be collected and maintained at any specific time. The membership of the DC and the KB of such communities will evolve and change, and must therefore be monitored over time.

The recursive nature of RI also allows different communities to use the Information Object in different ways. Depending on the relevant KB, the recursion will end at different points; the CURL Exemplars in Digital Archives (CEDARS) project referred to this concept as a Gödel end [15]. For example, a file format such as the Crystallography Information File (CIF) [16] could be readily understood by someone whose knowledge base includes the CIF file format. However, someone who has never heard of CIF would need additional RI, such as a software utility or the written CIF standard to make sense of the data.

The implicit and dynamic nature of knowledge means that a KB may be difficult to define; however, one possibility is to describe it as a set of familiar software applications, community standards, contextual descriptions and topic categorizations; research is in progress to identify and record a KB using more formal techniques [9].

3. Registry/Repository of Representation Information (RRoRI)

The Digital Curation Centre (DCC) [5] and the Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) Project [9] are developing a Registry/Repository of RI (RRoRI) [17]. This is intended to be an authoritative source of RI for those 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 regard to curation strategies. The work is heavily based on the ideas in the OAIS model; and centres on the notion that RI is critical to the long term access, understandability and usability of digital information.

The huge burden of collecting and maintaining adequate RI requires collaborative effort. Emphasis is therefore placed on interoperability and automated use, the vision being to establish a global, distributed network of RI which provides an infrastructure of reliable and trusted RI. Consequently, working with projects such as PRONOM [18] and the Global Digital Format Registry (GDFR) [19] is of particular relevance. These are subtypes of RI registries and focus on the provision of details about file formats (largely the structure type of RI) [20].

The current implementation of RRoRI [17] 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.

Access to RI by third parties is enabled through the use of two key concepts: Curation Persistent Identifiers (CPIDs) and descriptive RI labels [21] (see Figure 2 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. EPSRC National Crystallography Service

As part of the eBank-UK project, the EPSRC UK National Crystallography Service has constructed an institutional data repository (eCrystals [22]) 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 [23]. This work has paid particular attention to deriving a schema to describe the capture of data and contextual information throughout the entire workflow as opposed to purely the final result.

The data repository comprises a public and a private part; through the use of an embargo schema, data can be stored 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.

4.1 Crystal Structure Determination Workflow

Chemical crystallography is concerned with determining the structure of a molecule as well as its 3D orientation with respect to other molecules in a crystal. This is achieved through the 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 1. 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.

The primary aim of the repository is to make available and encourage the sharing of data, which are generated throughout the experiment pipeline shown in Figure 1.

During the work-up of the data, they progress from being in a state of raw to derived to final results data. The data collection stage provides JPEG files as representations of the raw data, which are derived from proprietary formats generated natively by the instrumentation used for the experiment. This stage may also have an HTML report file associated with it, providing information relating to machine calibrations and actions and as well as metadata describing how the data were processed. A significant result of the processing stage (process and correction of images) 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 (a suite of programs for crystal structure determination from single-crystal diffraction data).

Solving the structures 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). The SHELXL software produces both an output {.res} and a log file {.lst} in ASCII text format as a result of the data work-up process (this is an iterative refinement of many cycles and the output of the final stage is provided in the repository record). 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.

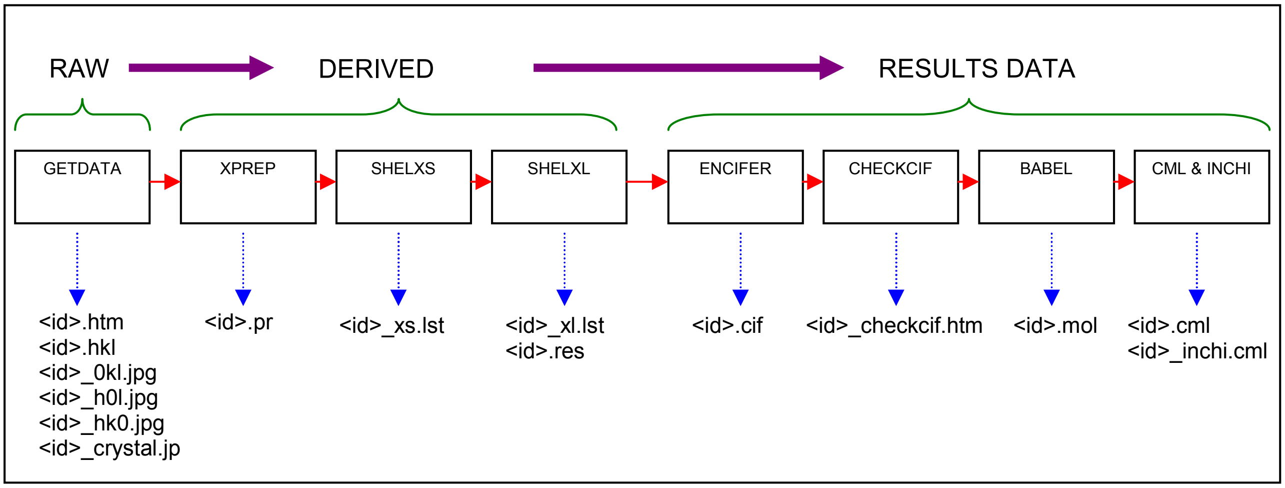


Figure 1: Workflow model of the EPSRC UK National Crystallographic Centre

The derived data are then converted to results data in the form of the CIF file format, which is used within the designated community as an interchange format and is supported by the International Union of Crystallographers (IUCr) – a publisher and learned society within the domain. CIF is a publishing format as well as being structured and machine-readable; it is capable of describing the whole experiment and related modeling 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 [24].

Another type of file format included in the final results data is a Chemical Markup Language (CML) encoding. 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, a unique text identifier that describes molecules, and is generated from the {.cif} file (note that the InChi can also be expressed as a URI in the info:inchi namespace). The file format conversions are performed according to well defined standards using the OpenBabel software obtainable from SourceForge.

5. 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 the 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 effectively 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. All of this type 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.

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 [25] providing an indication of the complexity and granularity of the information required. The RI Network in Figure 2 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.

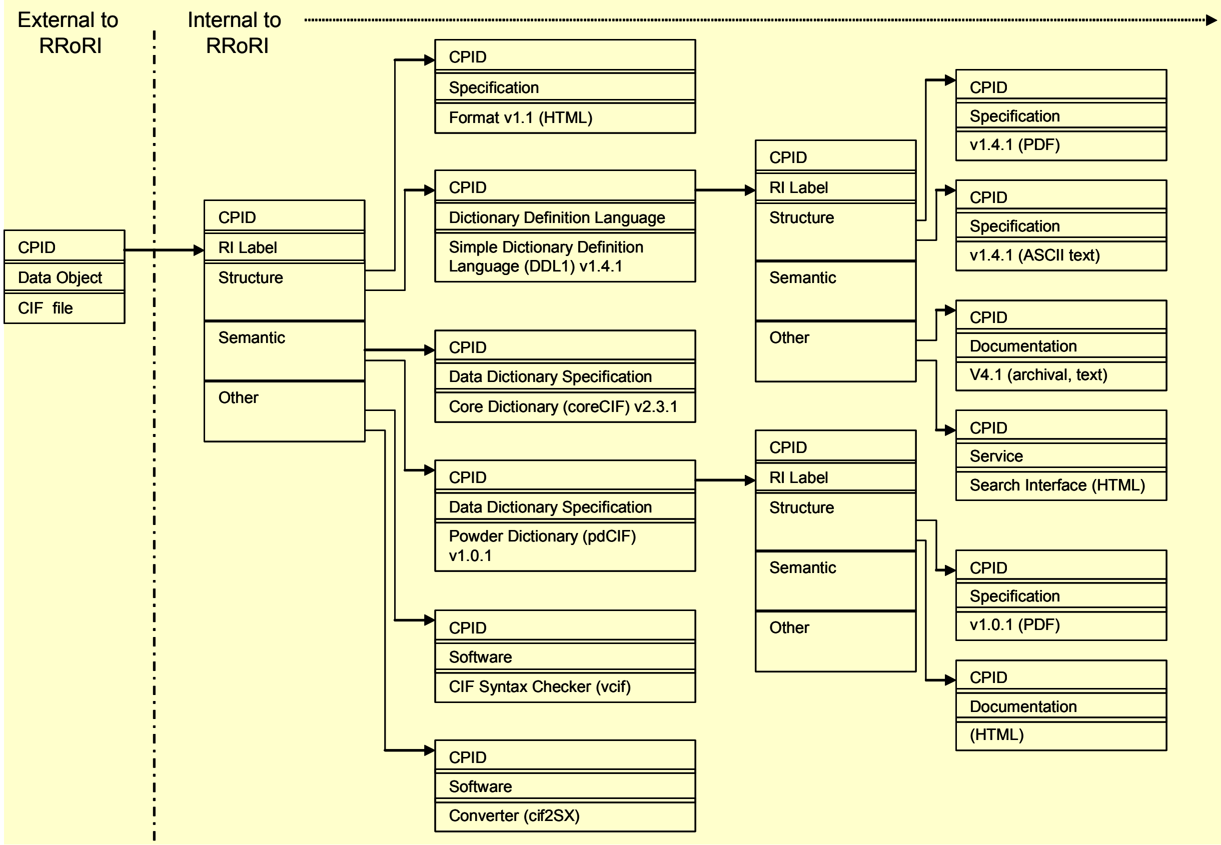
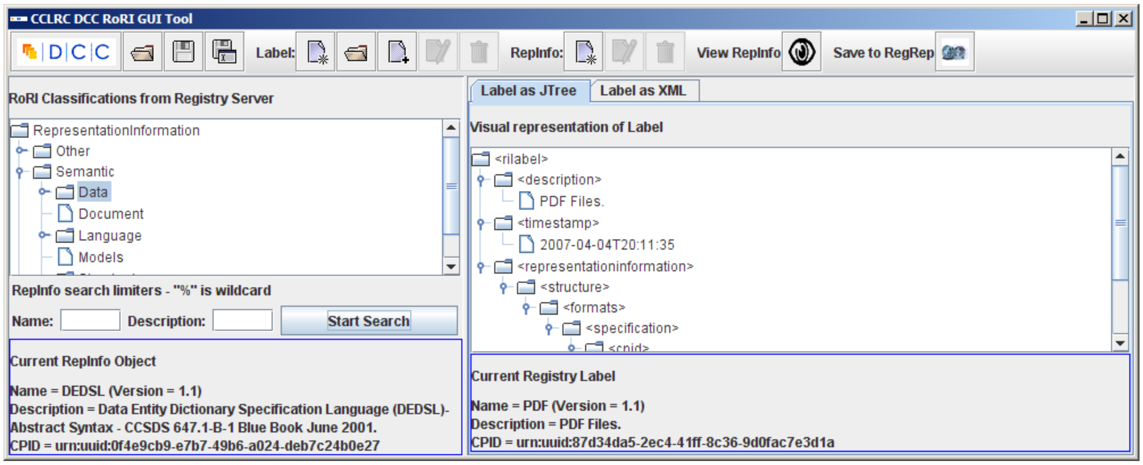


Figure 2: Graphical visualisation of part of an RI Network for the CIF file format [25].



RepInfo search limiters - "%" is wildcard

Name: Description:

Current RepInfo Object

Name = DEDSL (Version = 1.1)
Description = Data Entity Dictionary Specification Language (DEDSL)-Abstract Syntax - CCSDS 647.1-B-1 Blue Book June 2001.
CPID = urn:uuid:0f4e9cb9-e7b7-49b6-a024-deb7c24b0e27

Current Registry Label

Name = PDF (Version = 1.1)
Description = PDF Files.
CPID = urn:uuid:87d34da5-2ec4-41ff-8c36-9d0fac7e3d1a

Figure 3: A client with a GUI for ingest, search and retrieval of RI and RI labels.

6. 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 3). 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. An additional utility 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.

7. RI Network Usage Scenario

Most scientific data are now "born-digital" and rely 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.

8. Discussion and Further Work

A comprehensive discussion of issues relating to the use of RI in digital curation is provided in [26]. 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 4.1). 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. 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 complete 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 [6]: *Reference*: mechanisms to provide unambiguous access e.g. an object or a persistent identifier; *Provenance*: historical information to provide some assurance as to the likely reliability of the data; *Context*: relationships of the content information to other information e.g. calibration history; *Fixity*: data integrity checks to ensure authenticity e.g. encoding and error detection schemes such as checksums.

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. 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.

9. Conclusions

The prolific generation of digital data by disciplines such as crystallography necessitates curation and preservation. The OAIS concept of Representation Information (RI) is essential in identifying the dependencies involved in maintaining intelligible access to digital data over the long term. Collection and maintenance of suitable RI mitigates the difficulties related to the preservation of understandable information.

We have undertaken the first steps towards explicitly recording and maintaining RI for the CIF file format – a format that is vital to the recording and understanding of crystal structures both in the present and the future.

Acknowledgments. The Digital Curation Centre and the eCrystals Federation Project are funded by the UK's Joint Information Systems Committee. The CASPAR Project is an Integrated Project co-financed by the European Union within the Sixth Framework Programme.

References

[1] 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

[2] Hey, A.J.G., Trefethen, A.E.: The Data Deluge: An e-Science Perspective, in F. Berman, G.C. Fox and A.J.G. Hey Eds. Grid Computing-Making the Global Infrastructure a Reality, chapter 36, pages 809-824, Wiley & Sons. (2003)

[3] Carlson, S.: Lost in a Sea of Science Data, The Chronicle of Higher Education, (2006),

[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] The Digital Curation Centre (DCC), http://www.dcc.ac.uk/

[6] 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

[7] Strodl, S., Becker, C., Neumayer, R., Rauber, A.: How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure, Proceedings JCDL'07, Vancouver, British Columbia, Canada. (2007)

[8] PREMIS Data Dictionary for Preservation Metadata version 2.0, Preservation Metadata Maintenance Activity (2008), http://www.loc.gov/standards/premis/

[9] Giaretta, D.: The CASPAR Approach to Digital Preservation, International Journal of Digital Curation, Vol. 2 (1) (2007)

[10] Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Version 1.0, Center for Research Libraries and RLG Programs (2007)

[11] Tzitzikas, Y.: On Preserving the Intelligibility of Digital Objects through Dependency Management , Proceedings International Conference PV'2007, Oberpfaffenhofen/Munich, Germany (2007)

[12] The EAST Data Description Language, http://east.cnes.fr/english/index.html

[13] 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)

[14] Data Format Description Language, http://forge.gridforum.org/projects/dfdl-wg/

[15] CEDARS Guide to: The Distributed Digital Archiving Prototype, http://www.leeds.ac.uk/cedars/guideto/cdap/guidetocdap.pdf

[16] CIF -The Crystallographic Information File, http://www.iucr.org/iucr-top/cif/

[17] Registry/Repository of Representation Information (RRoRI), http://registry.dcc.ac.uk/

[18] PRONOM Registry of Technical Information, The National Archives, UK, http://www.nationalarchives.gov.uk/aboutapps/PRONOM/

[19] Global Digital File Format Registry (GDFR), http://www.gdfr.info/

[20] Brown, A.: PLANETS White Paper, Representation Information Registries (2008), http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

[21] DCC Development Team: DCC Label Report, http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCInfoLabelReport

[22] The Crystal Structure Report Archive –eCrystals Data Repository, http://ecrystals.chem.soton.ac.uk

[23] 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)

[24] IUCr checkCIF validation service, http://checkcif.iucr.org/

[25] Example Representation Information Network for CIF file format, http://homes.ukoln.ac.uk/~lismp/ECDL2009/RINetCIF.html

[26] 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/