Supporting e-Research Using Representation Information

David Giaretta[1,4], Manjula Patel[2,4], Adam Rusbridge[3,4],
Stephen Rankin[1,4], Brian McIlwrath[1,4]
[1] CCLRC, Rutherford Appleton Labs, Didcot, UK,
{D.L.Giaretta, S.E.Rankin, B.K.McIlwrath}@rl.ac.uk
[2] UKOLN, University of Bath, Bath, UK,
M.Patel@ukoln.ac.uk
[3] HATII, University of Glasgow, Glasgow, Scotland, UK
a.rusbridge@hatii.arts.gla.ac.uk
[4] UK Digital Curation Centre (http://www.dcc.ac.uk)

Abstract

Preservation has been defined as interoperability with the future. The OAIS Reference Model [1] identifies the role Representation Information plays in maintaining access to the information content of digital data. We discuss an additional role that Representation Information can play in facilitating interoperability to support e-Research on the GRID right now. This allows us to identify Representation Information as key to digital curation and a Registry Repository of Representation Information as an important service for preservation as well as e-Research.

1. Introduction

In preserving access to digital information over time, the importance of managing digital data in a well defined and structured archive has increasingly become well accepted. However, passive retention is not enough: digital objects inherently require additional information and methods to convert data into a form that can be interpreted for use and reuse. The OAIS reference model [1] identifies this additional information as Representation Information. Unfamiliarity with data in formats that are obsolete can present the same problems as unfamiliar data from another discipline. Overcoming this unfamiliarity has been complicated by rapid technological changes that in many cases results in the unavailability of software and documentation. Developing a structured registry and repository to store and maintain this Representation Information in a form that assists long-term use in conjunction with GRID related technologies may benefit the use and interoperability of contemporaneous data.

2. Representation Information

Representation Information (RI) is defined as the information required to allow a digital object to be converted to an information object. As shown in Figure 1, items of RI can be categorised to describe different roles. The following paragraphs discuss these classifications in further detail.

Structure Information describes and imposes restrictions on the internal structural composition and relationships of a data object. This may be based on a standard or specification, but whereas those define the legal elements and associated data types they tend not to impose restrictions on their usage and so further details are often necessary (consider here the distinction between the HTML specification and the HTML Strict DTD). Structure Information could be simple text but to facilitate automated processing more formal descriptions of file formats are advantageous, for example in EAST [2], FLAVOR [3], or DFDL [4] languages. Structure Information is not necessarily available at the time of creation, and so additional effort may be required to generate it. Some digital objects, such as those created by proprietary software, can also have unknown structure. In this case, the original software, or some equivalent application, may be required to enable access.

Semantic Information 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. Identifying and describing semantic information in the most general terms remains a difficult problem and is a topic of active research. Current work focuses on data dictionaries, and ontologies.

Other RI may indicate software, standards, algorithms, packaging information, and documentation that relate to the digital object.

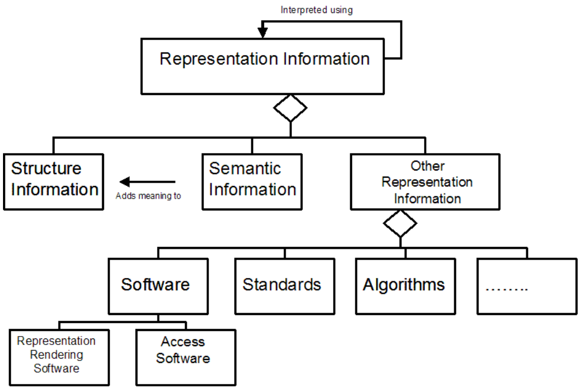

Figure 1 RI Network (extended) [1]

In many cases, the RI applicable to a data object is itself a digital object and may itself require further RI to enable it for use. In this way, a recursive Representation Network is created. Without careful management, this Network can scale poorly and require an infeasible quantity of RI to be collected. A limit may be placed on the amount of RI required at any one time by using the concept of a Designated Community and its Knowledge Base. A Designated Community is a group of users for whom the data is being maintained, and its Knowledge Base is the level of knowledge and software applications with which they are familiar. The membership and knowledgebase of these communities will change over time, and there will also be technological changes that take place. As a result of these changes the RI required will change.

By sharing the collection of RI between organisations, the effort required may be reduced, and the feasibility of collection increased. The Designated Community will have to be monitored in some way to ensure the Representation Information continues to be suitable for its needs. As an example of Representation Information, consider an astronomical image in FITS file format [5]. This contains a header with keyword value pairs, where each keyword has a specific meaning, although only a small number of these keywords are standardised. The Semantic Information should explain the context of each keyword.

The file in FITS format would be easily understood by someone who knew how to handle this format, someone who's Knowledge Base includes FITS (for example, they have some appropriate software). Without this knowledge, additional RI would be needed. Additional RI will take the form of software to display the image (which may need additional RI in the form of an emulator) or a human-readable standard describing the FITS file format.

2.1 The Long Term Role of Representation Information

Without maintaining and retaining suitable RI, the difficulties of preservation of information, as opposed to bytes, are increased. Data formats, software, standards, and programming languages become obsolete; the documentation for these are often poor or non-existent; and the specialised knowledge needed to manipulate these is often not externalised.

Submitters of data to an archive must be able to identify the RI needed by the end users. Policies must be developed specifying how suitable and comprehensive RI. Although RI allows the formal identification of the composition of a digital object and indicates an environment in which the data was previously used, it may be difficult to encourage submitters of data to provide corresponding RI. Licensing restrictions may limit the availability of some essential objects. The additional step of RI collection in the workflow process may restrictively increase the cost of archiving. In order to overcome these issues, clear policies of retention and efforts towards workflow streamlining are needed.

2.2 Contemporaneous use of Representation Information

RI by definition is what is needed to facilitate use and reuse of digital information. The RI may be limited, but adequate for a contemporary Designated Community. However for contemporary users outside the Designated Community, this RI may be sufficient to obtain further contemporary resources. Alternatively additional RI may be provided to support this broader Community. In many instances, this reuse may entail the mapping of information between structures to facilitate interoperability. Between, and even within disciplines, these methods are often not shared through a dedicated infrastructure. Sharing this RI increases access to resources and reduces the cost involved in transforming data into a structure suitable for specific purposes.

3. The DCC Representation Information Registry and Repository

In order to help to avoid duplication, share resources, coordinate access, and minimise effort, a dedicated and well engineered network of Representation Information Registry-Repositories (RI-RRs) are needed. The DCC is actively developing a Representation Information Registry (DCC- RR). This is not intended to be a data repository, but an authoritative source of RI for the community responsible for the collection, curation, and management of data.

We see our primary function as providing and sharing information that enables managers of digital information to make informed decisions with regards to curation strategies. The DCC-RR aims to make relevant RI available in a readily accessible manner to enable third parties to make informed decisions with regard to the management of their data [6].

An early prototype Representation Information Registry (DCC-RR) is currently available, intended as a proof-of-concept demonstrator [7]. The ideas behind this approach are detailed in the document, DCC Development Approach to Digital Curation [8]. At present the prototype DCC-RR caters for: viewing RI already registered; registering of new RI; creation of a new RI Label; and adding a classification entry.

We plan to work with projects such as PRONOM, developed by the National Archives, and the Global Digital Format Registry (GDFR), developed by the Digital Library Federation (DLF). These focus on the provision of details about formats – essentially limited to the Structure type of RI. These projects will help to inform the development of the DCC-RR, and strategies are being developed to ensure the data contained within the DCC-RR is interoperable with these.

3.1 Representation Information Labels

RI should be stored in a well defined and managed Registry/Repository (RI-RR) system, with appropriate metadata describing it.

A digital object should be associated with a structured label as a necessary (but not sufficient) condition for long-term preservation, through a type of logical attachment or packaging. This label, such as that described in [9], allows the identification and categorisation of RI and provides a set of entry points into the DCC-RR. As each RI object has an associated label identifying suitable rendering conditions, a directed Representation Net can be created.

For long-term preservation each entry of RI should be identified with a persistent identifier. A label, which can be attached to either a data object in an external repository or an RI object internally in the RI-RR, then contains a set of identifiers. This label then serves as a set of entry points into the RI-RR, allowing the Representation Network to be traversed and the appropriate RI discovered. Ideally, this operation will be automated and transparent to the end user. The variety of formats, software, specialised subjects, and user communities that exist indicates that a distributed network of Registries may be needed. We hope that any foundations laid now may develop into a distributed, global, and federated collaborative system.

In the long term, we hope to develop identification services for the DCC-RR which allow a user to submit an object which is then identified and an appropriate label returned.

4. Interoperability and Automated Processing

The interconnected nature of the Representation Network and the complexity of software mean that where possible automated processing should be emphasised. This scenario is clearer when considering science formats, many of which are designed for automated processing by a parsing program. To handle an "unfamiliar bit stream" without using specially written software one would need a way to describe data at different levels of abstraction, for example from the bit level to complete information objects (see Figure 2), for use in general purpose applications and interfaces which will support interoperability.

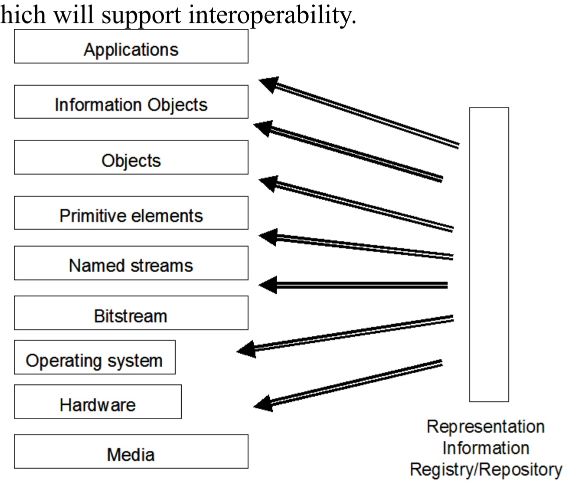

Figure 2 Role of Representation Information

We are investigating the use of data description languages such as EAST [2] and DFDL [4] to describe structure information in detail, and how these may be used to map between alternative formats.

The EAST language is well supported with tools to aid the production of EAST descriptions and associated Data Dictionaries. Accordingly, we are in the process of integrating EAST tools into the DCC-RR to enable us to get from bit level representations to primitive objects.

5. Web Services

Current development of the DCC-RR suggests that a network of geographically distributed and federated repositories will be able to contain the greatest level of knowledge, providing the most complete level of support. Access to RI should facilitate the use and re-use of digital information across disciplines, as is increasingly likely in e-Research. In particular we are developing a number of generic applications and APIs which will allow one to perform useful operations on digital information from unfamiliar domains, in unfamiliar formats, given adequate RI. A particular type of such an application can be used to provide services such as format transformation. These will be available as Web Services, and we further hope to collaborate with and build upon the work of the PANIC [10] project.

6. Use of Representation Information in the GRID

The use of the GRID and e-Research in general will make use of "unfamiliar bit streams" in the form of data relevant to a piece of research but perhaps not in a familiar discipline. It is important that this data is interpreted correctly by the user. RI provides access to methods that will further the ability of users to interpret this data. In addition, in order to support processing of potentially large amounts of such information with tools familiar to the user, as far as possible the RI should make the data usable by the user's general purpose software.

7. Benefits for e-Research

Concerns regarding information preservation are growing, particularly within the field of e-Science where the costs involved in data production are significant. The methods by which information is interpreted are likely to change over time, and so maintaining a structured repository record of RI ensures previously methods can be accessed and viewed for strengths and weaknesses, allowing the most appropriate use, or recreation, of some particular functionality. By developing an environment in which data descriptions can operate in conjunction with web services and GRID applications, the ability to share and automate the use of information is facilitated.

8. Conclusions

RI is a key idea in managing and using distributed digital data as well as preserving digital information over the long-term. The UK Digital Curation Centre is working on this basis for a globally distributed network of Registry/Repository for Representation Information which will support both these types of activities. Key concepts relating to interoperability, automated use, and web services have been identified, and development work is ongoing to implement these. These ideas show the dual use to which RI can be put, namely preservation activities as well as e-Research using contemporaneous data.

References

1. ISO 14721, Reference Model for an Open Archival Information Model, http://www.ccsds.org/documents/650x0b1.pdf
2. EAST, ISO 15889, http://east.cnes.fr/
3. FLAVOR, http://flavor.sourceforge.net
4. DFDL, http://forge.gridforum.org/projects/dfdl-wg/
5. FITS Format, http://fits.gsfc.nasa.gov
6. DCC Development Team Report, Scoping Study for a Standards Registry, http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCStandardsRegistry
7. DCC Development Team, DCC-RR Development, http://dev.dcc.rl.ac.uk/dccrrt/
8. DCC Development Team, DCC Approach to Digital Curation, http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCApproachToCuration
9. DCC Development Team, DCC Label Report, http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCInfoLabelReport
10. Semi-Automated Preservation and Archival of Scientific Data using Semantic Grid Services; Hunter, J. and Choudhury, S. May, 2005, http://metadata.net/newmedia/Papers/SIGAW2005_paper.pdf