Curating for the Future – The work of the Digital Curation Centre

David Giaretta, Liz Lyon, Bridget Robinson

The Digital Curation Centre

Abstract

The poster will present the background to the establishment of a Digital Curation Centre and will chart the development of the DCC in the initial set-up period Phase One through to the plans for the formal launch of the DCC in November 2004. It will include details of the structure of the DCC, its overall mission, aims, objectives, and early deliverables. These will include descriptions of the research plans and development activities, the advisory service and the outreach functions. Plans for the longer term will also be outlined.

1. Background

Scientists and researchers across the UK are generating large amounts of data through experimental methodologies, simulations and observational techniques. This volume is increasing dramatically with the advent of Grid-enabled applications. The e-Science Data Curation Report[1] highlighted the importance of ensuring that this deluge of data being created in e-science and e-research should remain available and valid for future researchers. (The findings of the report will be covered in the paper " From Data Deluge to Data Creation"[2]). In response to these findings the JISC (Joint Information Systems Committee) and the e-Science Core Programme (e-SCP) put out a call in 2003 for bids to establish a National Digital Curation Centre. The purpose of the centre would be to lead research and development into key areas of data curation and preservation and to pilot the development of generic support services for maintaining digital data and research results over their entire life-cycle for current and future users.

2. The Digital Curation Centre

With a start date of March 2004, JISC appointed a consortium to set up and run the new Digital Curation Centre for three years providing development, advisory and outreach services to the wider community. e-SCP research funding will begin on 1 September 2004. The DCC is based at the National e-Science Centre (NeSC) jointly managed by the Universities of Edinburgh and Glasgow. The formal launch of the DCC will be in November 2004.
The consortium is made up of four partner institutions.

- University of Edinburgh (Informatics, Law, Information Services and leading research institutes)
- University of Glasgow (HATII and Information Services)
- UKOLN, at the University of Bath
- The Council for the Central Laboratory of the Research Councils (CCLRC)

3. Aims and Objectives

The overall objective of the DCC is to provide a centre of expertise in data curation and preservation. It will be driven by the twin aims of excellence in research and excellence in service and will seek to gain international respect and national leadership.
In order to meet these aims the DCC will concentrate its work in the following areas:

3.1 Research

- To draw together the functions of curation from traditional archival methods to the curation of evolving knowledge as seen in scientific databases
- To identify through direct collaboration the key fields in which research is needed
- To conduct research in areas already identified as crucial to data curation e.g. annotation, data integration and publication, appraisal and long-term preservation
- To link research and services so that practical issues can be addressed by researchers and products of research can be tested in practice

3.2 Development

- Offer a repository of tools and technical information
- Develop services to evaluate tools, methods, standards and policies
- Registry of metadata standards
- Registry of file formats

3.3 Advisory Services

- Curation Manual for practitioners
- Current Awareness Service
- Technical Standards Framework
- Helpdesk **digitalcuration@ed.ac.uk**

3.4 Community Support and Outreach

- Website **http://www.dcc.ac.uk**
- DCC e-journal
- Training and Professional Development

Figure 1 shows the interaction between these areas, which we believe will form a "virtuous circle".

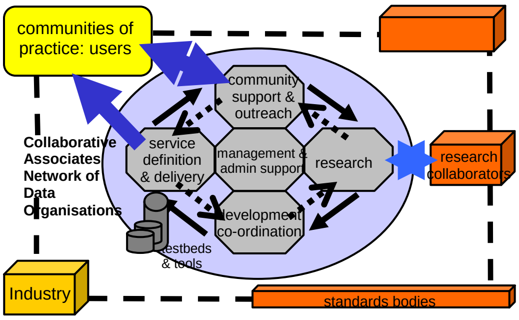
communities of practice: users
community support & outreach
service definition & delivery
management & admin support
research
research collaborators
development co-ordination
testbeds & tools
Collaborative Associates Network of Data Organisations
Industry
standards bodies

Figure 1

4. User Requirements Analysis

The analysis will set the contextual background, examining the work that is being done by large-scale e-science and e-research initiatives and identifying the data curation issues that they face.
In the first phase the analysis will concentrate on data creation, data curation and data re-use communities. It will consult representative users from:

- science, social science and humanities research
- the learning community
- relevant industry and commercial organisations
- the library and information science community

In the second phase the objectives will be:

- to align user requirements with the goals and objectives of policy makers
- to increase the understanding of digital curation user requirements within the DCC and across a range of relevant policy-making organisations

5. The Associates Network

The DCC is creating an Associates Network drawing on prominent members of three communities:

- UK data organisations
- leading data curators overseas and supranational standards agencies
- industrial and business communities

The purpose of the Associates Network is to ensure that the DCC addresses the needs of representatives from all the areas of digital curation. The Associates will provide:

- input to steer the research programme
- national and international collaboration in the development of standards, methodology and best practice
- consistency between standards for digital curation defined within the academic and commercial worlds, and a sharing of expertise

6. The DCC Approach to Digital Curation

Underpinning the work already outlined, the DCC will be developing an overarching "Approach to Curation" which will include a conceptual model and an outline architecture. The OAIS Reference model (**http://www.ccsds.org/CCSDS/documents/650x0b1.pdf**) and its view of information preservation will be used as a basis for this approach. Ideas from other sources, which are not within the remit of OAIS, will be added. In particular we will address issues of "curation" which go beyond the "digital information preservation" covered by the OAIS. We also bring in ideas from the e-Science community about distributed use of data, automated processes, interoperability, and data virtualisation.

In developing this architecture the DCC will work closely with the Global Grid Forum (GGF) and the Persistent Archives Working Group (PAWG) (http://www.zib.de/ggf/data/pa/charter.html) which is looking at looking at persistent archives, which are defined as providing the mechanisms needed to manage technology evolution while preserving records and their context, and basing these on virtual data grid technology, focussing on *"the management of the evolution of the software and hardware infrastructure over time"*. Persistent archives may be viewed within the context of the OAIS Reference Model, covering bit preservation, migration, media renewal etc, but omitting the designated community and many aspects of preservation planning. Nevertheless it is clear that the Data Grid technology will be a very valuable area of work which the DCC should be able to use, and perhaps contribute to.

Other preservation architectures are being developed, notably the National Digital Information Infrastructure Preservation Programme (NDIIPP) (http://www.digitalpreservation.gov/index.php?nav=3&subnav=12) from the US Library of Congress.

Within this broad architecture we need to identify a clear niche for ourselves, as well as identifying partners with whom to work. The DCC will put in place a number of services early on, such as the advisory service and helpdesk. In addition to these we will provide initial implementations of substantive services to support automated and distributed data re-use – and preservation. As an example of this approach one can consider the Representation Information Repository.

7. DCC and File Format Registries

Within the call to set up the DCC there was a requirement that a Formats Registry be created.

There are already several efforts in this area, including PRONOM[3] and the Global Data Format Registry[4]. However the DCC is approaching this as follows:

- We will, in the spirit of OAIS, generalise the File Format concept to “Representation Information”
- We will initially look at science formats since other registries focus on document formats (i.e. we identify our niche)
- We will try where possible to facilitate the use of tools to allow re-use of the data – in a way that is independent of existing software (to support the re-use aspects of Curation)
- We will design the system such that it could be a component in a distributed application (to support e-Science)
- We will aim to work in partnership with existing Format registries to provide a service that is robust and covers as many formats as possible.

Looking at these points in more detail, we have the UML diagram for the OAIS Information Object

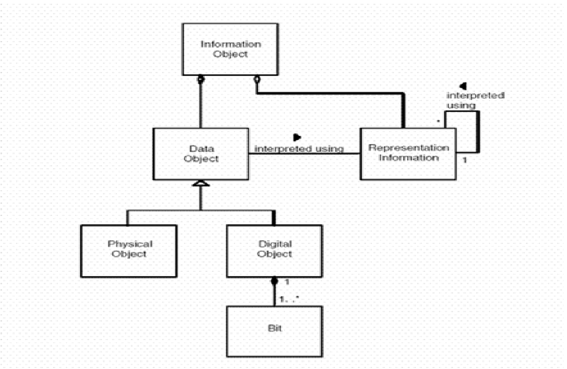

Figure 2

Figure 2 shows that (a) Representation Information is needed to convert data into information and (b) Representation Information may itself need further Representation Information in order to be understood.

To see where Formats come in this scheme the Figure 3 shows that Representation Information has a part called “Structure” – this contains the format information.

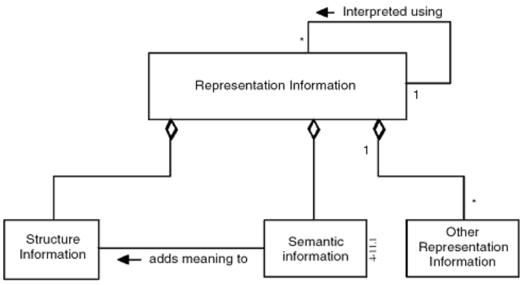

Figure 3

Clearly there are also other components – this is where we generalise the Format Registry and aim to contain Semantic information (e.g. Dictionaries or Ontologies and more), plus any other type, for example software, which may be needed.

There are already tools such as the EAST[5] suite, which allow one to describe data down to the bit-level and then use that description to extract information from the bit stream using generic tools. Clearly in the long term this suite of tools may fall out of use however, since EAST is an ISO standard, we assume that the standard itself will be obtainable and the tools re-implemented.

Thus the Representation Information will form an important component in Long-Term Preservation of digital information. At the same time it will facilitate the use of generic tools. The link between the Representation Information and the data will be via DCC persistent identifiers (shown as RILabel in

Figure 4). Preservation Planning (c.f. OAIS) will be helped by extension of the Representation Net within the DCC.

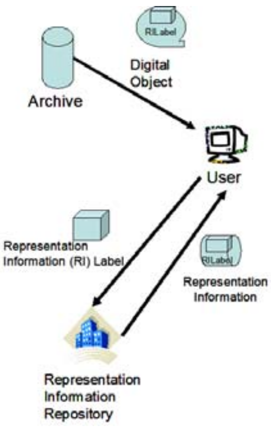
RILabel
Digital Object
Archive
User
Representation Information (RI) Label
RILabel
Representation Information
Representation Information Repository

Figure 4

For example consider data that is stored as a compressed FITS file. The identifier will give access to information about the compression as well as the FITS format via many protocols including for example Web Services. Rather than simply point to human-readable documentation, we may point initially to software that can be used. When that software is no longer usable the DCC will instead preferentially offer bit-level descriptions, which are usable by generic tools.

This mechanism of redirection and update will lift some of the burden of tracking change from digital repositories, many of which currently duplicate work many times over.

7.1 Time-varying and Reactive Systems

Worthy of special attention are the many instances, (some would argue these are in fact the majority), where the data is changing over time, in particular databases. Significant research efforts are being applied to this problem.

In addition support will be provided for those data types to which access is available only through special software.

8. Conclusions

The DCC should provide unique support for Digital Curation in the UK. Furthermore it is well placed to be a global focus for expertise in this area which is relatively young but rapidly growing in importance both in the academic as well as the commercial worlds.

References

[1] Formally published by JISC. See Data Curation for e-Science in the UK. Lord, Philip & MacDonald, Alison. 2004

[2] From Data Deluge to Data Curation, Philip Lord, Alison MacDonald & Liz Lyon. E-Science All Hands Meeting 2004

[3] http://www.records.pro.gov.uk/pronom/

[4] http://hul.harvard.edu/gdfr/

[5] http://east.cnes.fr/