The Digital Curation Centre: A Vision for Digital Curation

Chris Rusbridge

Digital Curation Centre (DCC), University of Edinburgh
c.rusbridge@ed.ac.uk

Peter Burnhill

Digital Curation Centre (DCC), & EDINA, University of Edinburgh
p.burnhill@ed.ac.uk

Seamus Ross

Digital Curation Centre (DCC), & HATII, University of Glasgow

s.ross@hatii.arts.gla.ac.uk

Peter Buneman

Digital Curation Centre (DCC) & School of Infomatics
University of Edinburgh

David Giaretta

Digital Curation Centre (DCC), & Council for the Central Laboratory of the Research Councils (CCLRC)
d.l.giaretta@rl.ac.uk

Liz Lyon

Digital Curation Centre (DCC), & UKOLN

e.lyon@ukoln.ac.uk

Malcolm Atkinson

National e-Science Centre,
Department of Computing Science, University of Glasgow
&
School of Informatics, University of Edinburgh
mpa@nesc.ac.uk

Paper For:

From Local to Global: Data Interoperability--Challenges and Technologies,

Mass Storage and Systems Technology Committee

of the IEEE Computer Society,

20-24 June 2005, Sardinia, Italy

The Digital Curation Centre: A Vision for Digital Curation

Chris Rusbridge
Digital Curation Centre (DCC), University of Edinburgh
c.rusbridge@ed.ac.uk

Peter Burnhill
Digital Curation Centre (DCC), & EDINA, University of Edinburgh
p.burnhill@ed.ac.uk

Seamus Ross
Digital Curation Centre (DCC), & HATII, University of Glasgow
s.ross@hatii.arts.gla.ac.uk

Peter Buneman
Digital Curation Centre (DCC) & School of Infomatics University of Edinburgh

David Giaretta
Digital Curation Centre (DCC), & Council for the Central Laboratory of the Research Councils (CCLRC)
d.l.giaretta@rl.ac.uk

Liz Lyon
Digital Curation Centre (DCC), & UKOLN
e.lyon@ukoln.ac.uk

Malcolm Atkinson
National e-Science Centre, Department of Computing Science, University of Glasgow & School of Informatics, University of Edinburgh
mpa@nesc.ac.uk

Abstract

We describe the aims and aspirations for the Digital Curation Centre (DCC), the UK response to the realisation that digital information is both essential and fragile. We recognise the equivalence of preservation as "interoperability with the future", asserting that digital curation is concerned with 'communication across time'. We see the DCC as having relevance for present day data curation and for continuing data access for generations to come. We describe the structure and plans of the DCC, designed to support these aspirations and based on a view of world class research being developed into curation services, all of which are underpinned by outreach to the broadest community.

Introduction

The foundation of the Digital Curation Centre (DCC) reflects the belief that long term stewardship of digital assets is the responsibility of everyone in the digital information value chain [15]. The maintenance, usability and survival of digital resources depends on regular planned interventions; care needs to be taken at conception, at creation, during use, and as use transitions to lower levels. It may be tempting to view e-Science or Cyber-infrastructure as concerning huge, often distributed databases in current use, quite distinct from viewing digital preservation as concerned with more common, much smaller files, ensuring they are available for the future. The *Open Archival Information System* (OAIS) model's [5] emphasis on ingest, followed by preservation management, and then later acts of dissemination, does tend to suggest a model where resources are put away in safe keeping for the future. In reality, the DCC views digital curation as a continuum of activities, supporting the requirements for both current and future use [21]. Huge datasets in current use need action now to ensure their future utility, while digital preservation has always deprecated 'dark' inaccessible archives.

The long term value of data rests in their potential as evidence, their reuse possibilities, and their role in facilitating compliance and in ameliorating risk. Curation, the active management and care of data is the key to realising that potential. As scholarly research and scientific study becomes increasingly driven by the analysis of data, long term access to these data is

crucial in enabling the verification of scientific discovery and to providing a data platform for future research.

This view of curation embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage: placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration. Context and linkage are terms rich in intention, with implications for metadata and interoperability. Resource discovery and retrieval requires mark-up with time/place referencing as well as subject description and linkage to discipline-based ontology.

Only by promoting the ideas that underpin digital curation from the conception and creation of our digital assets until long after they have passed out of their primary usefulness can we claim to have succeeded. The Digital Curation Centre, through its organisation, emphases and practical activities closely reflects these ideals and it aims to catalyse action in innovative research, development, service delivery, and outreach. Its primary aims are:

- to promote an understanding of the need for digital curation among the communities of scientists and scholars;
- to provide services to facilitate digital curation;
- to share knowledge of digital curation among the many disciplines for which it is essential;
- to develop technology in support of digital curation; and,
- to conduct long-term research into all aspects of digital curation

The DCC has been funded for three years in the first instance by the UK Joint Information Systems Committee (JISC) [24] and the UK e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC) [19]. Three quarters of the project's funding was awarded by the JISC from March 2004 with the remaining 25% coming on stream at the beginning of September later that year. By working with other practitioners and agencies the DCC will support UK institutions to store, manage and preserve digital assets to ensure their enhancement and their usability over the long term [26]. The DCC is the national focus for research into curation issues and to promote expertise and good practice, both nationally and internationally, in the management of all research outputs in digital format. The drivers which led to the establishment of the DCC are an increasing awareness within the scientific and research community that digital assets are reusable, that access to them is essential if contemporary scholarship is to be verifiable and reproducible, and that for many reasons ranging from degradation of media to metadata drift and to technological obsolescence they are fragile.

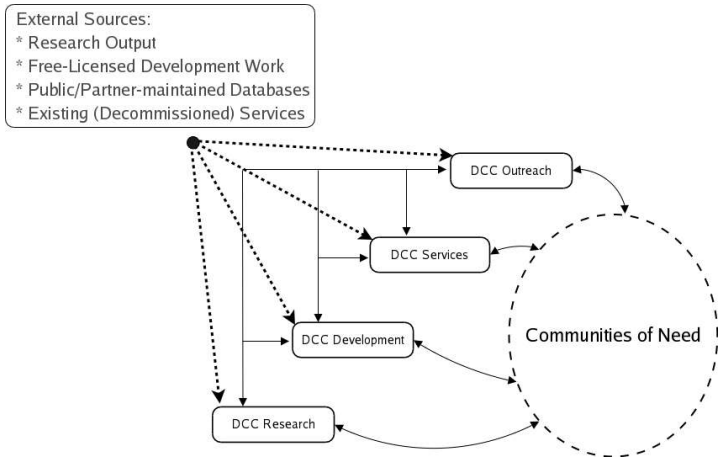
Engaging the community

The DCC represents a collaborative effort led by a consortium of four institutions, each bringing diverse experience in a range of Digital Curation areas. It brings together organisations across three Universities and a research council. Led by the University of Edinburgh [35], which hosts the School of Informatics [7,31], the National eScience Centre (NeSC) [27], the EDINA national data centre [17], the AHRC Centre for the Studies in Intellectual Property and Technology Law [2], the DCC consortium includes HATII [22] at the University of Glasgow [36], UKOLN [32] at the University of Bath [34], and the Council for the Central Laboratory of the Research Councils (CCLRC) [6]. Given the broadness and pervasiveness of the digital curation challenge the core partners recognise that a sustainable contribution can only be made if widespread activity can be leveraged. To ensure that this happens the partners have established mechanisms to support the development of a network of associates.

Engaging the community and creating conduits between our different activities with other research, development, services and user communities is essential if the DCC is to deliver the catalytic impact its funders envisaged in the call for expression of interest in JISC Circular 6/03, issued in September 2003 [25]. It is crucial for the Centre's success that it benefits from the diversity of expertise available within the UK digital curation community, and that the broadest range of communities needing digital curation support and services are engaged in its activities. These different communities should help steer the DCC's research programme, contribute to its community building events, and benefit from the advisory services that it operates. Science and scholarship are international in nature, and best practice in digital curation must reflect that. This is immediately obvious in such fields as astronomy, physics and bioinformatics, where data are increasingly created and curated by multi-national consortia. This requires international collaboration for shared development and consensus on methodology and standards. The demand for digital preservation technologies comes from and will be met by a wider community than academia; therefore, the DCC must engage with the industrial and business communities. In building such relationships the DCC will create outlets for the technologies and methods it develops and benefit from access to commercial developments.

DCC activity domains

The DCC activities have been separated into four key task areas, with an umbrella management group overseeing and coordinating the work of each. The four main areas are Research, Development, Services, and Outreach. Each activity domain has its own work programme. Many elements of these programmes can be delivered independently of the work being done in other domains, but the DCC will have proven itself most successful if it proves possible for us to move from concept through research and development to services and outreach. We hope that the outputs of individual task areas will benefit from inputs from contributions by members of the Associates Network. These contributions may come as the sharing of expertise, donations of technology, methodologies, or applications, and collaboration in research, development and service provision. Inter-task-area relationships are facilitated by the DCC's distributed nature, with representatives of each activity domain distributed across the four partner institutions. The communication infrastructure required to enable such collaboration and interaction is well established, and the Consortium relies upon a range of technologies (e.g. online fora, email, conference calls, and the Access Grid [1]) to allow the straightforward exchange of ideas and results. By driving our activities from a solid research footing into innovative and service-led development work we can meet our numerous service objectives and ensure an effective and dynamic outreach programme. Since the relationships between activities are non-linear and bidirectional in nature every task area can benefit from experiences acquired throughout the project's full extent and beyond.



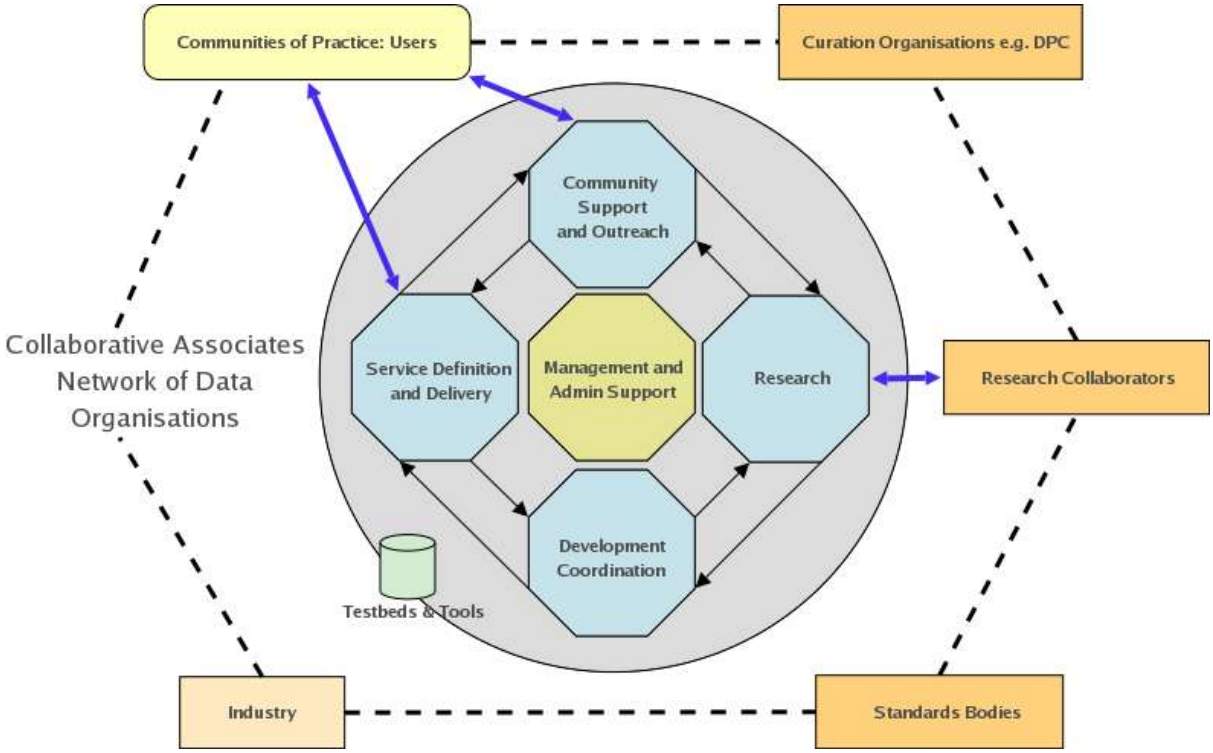
Task Area Inter-relationships

Curation is the key

Since its conception, the DCC has assumed a structure representative of an overall vision of Digital Curation. In order for one's digital curation endeavours to be successful, and consequently for a Digital Curation Centre to be successful, it is essential that a set of core activities are undertaken. The task area infrastructure and their associated work areas are designed to facilitate the completion of these activities. The DCC is perceived as a potentially long term project. Research being undertaken now might not come into fruition for some years, by which time it will be possible to trace a clear development, service and outreach path through which to disseminate the results. The DCC has a key role to play in developing and promulgating the changes in community practice that are needed if our research communities are to avoid the decay of our digital research and scholarship heritage.

Digital Curation itself is the active management of data over the life-cycle of scholarly and scientific interest; it is the key to reproducibility and re-use. Metadata for resource discovery and retrieval are important, with mark-up on time/place referencing as well as subject description and linkage to discipline-based ontologies providing key research foci. Special emphasis is required on the descriptive information that allows effective re-analysis of datasets of scientific and scholarly significance, and re-use in new and unexpected contexts, e.g. e-Learning or history of science. The demands for linkage to the two further domains of scholarly communication and e-Learning must also be understood. The Centre is pluralistic: it values and works with the many and diverse cultures that span the scholarly and scientific communities, and it will seek to value and understand the different paradigms and methodologies. It aims to address both generic and disciplinary perspectives.

Science and scholarship, more generally, cut across disciplinary boundaries. So too does digital curation: appreciation of differences between disciplines is an essential aspect for understanding and consensus building. An open and creative culture is necessary to foster the flow of ideas between research and practice, provoking research with practical challenges and informing practice from theory and experiment. This lays the grounds for leadership and advocacy, for continuing professional development, for matters requiring mediation, for building consensus, and for successfully enabling the adoption of common standards. Use of open standards enables considerably the re-use and re-purposing by avoiding dependencies



of platform and application software.

Fundamental to the success of digital curation is an understanding of the ubiquity of its relationship with digital information. Even before digital content is conceived, and even after it has fulfilled its primary usefulness, digital curation activities must be undertaken and promoted if digital materials are to remain viable. Curation is not a box to be ticked on a manifest or a single process through which data passes. It is an ubiquitous endeavour that should characterise all other interactions with and manipulations of digital content.

The necessity to share access to digital research resources has long been a driver within the sciences but pressure for action is now keenest within the physical and life sciences. As e-Science has become increasingly data driven such problems as scale, complexity, distribution, provenance and integration of data have become increasingly prominent. The 'document' (library and archive) tradition has an even longer claim on shared access, now augmented by pervasive internet access to digital objects; a foundation for global collaboration in the provision and use of digital resources as evidence in research. The DCC will evaluate and select best practices adopted by data libraries, data archives and data centres, promoting those it finds most effective and efficient. It will draw upon and enhance well-tested principles developed by the research library, records management and archive traditions as well.

Research

The DCC is research led. It has three main goals:

- To draw together the various functions of curation, from the traditional archival functions to the maintenance and publication of evolving knowledge as seen in scientific databases.
- To conduct research in areas already identified by the partners as crucial to digital curation.
- To institute two-way conduits between research and service in which real problems can be drawn to the attention of researchers and the products of research can be tested in practice.

A range of topics has been chosen, reflecting these goals. Each has in common the fundamental nature of their implicit problems, offering a likelihood that the research generated will be highly visible. In addition, every topic within the agenda will yield results that can be immediately exploited and investigated further through the subsequent work by the development and services teams. Finally, each of the initial research topics is relevant to the partner institutions and a broad cross section of the network of associates[9]. A research committee meets regularly with the task of

identifying emerging research opportunities. This committee is outward-looking, and partners and members of the network of associates feed ideas and needs into its discussions. This should keep our research vital and relevant.

There is much preservation research going on outside the UK with which we are engaged. There are researchers from the industrial sector who would benefit from taking time out and participating in the DCC research programme. These individuals bring fresh perspectives and create a conduit between the research laboratories and the DCC research community. To enable collaboration and communication with these communities the DCC has developed a visiting scholars programme.

The Research Agenda currently focuses on:

Data integration and publishing

Special emphasis is given here on integration techniques in the context of digital preservation metadata. Current work investigates publishing curated databases or parts of curated databases that other researchers or scientists may bring into or integrate with their own "research databases". This includes developing techniques to pull records out of differently structured relational databases and translate them into XML documents that share the same structure and can be more easily integrated.

Annotation

Annotation plays a fundamental role in scientific and scholarly work. Researchers use annotation as a mechanism to enhance the description or interpretation from trusted sources that may inform further interpretation of data drawn or research with different communities wishing to annotate different types of digital data (structured text, audio, images, video) in a variety of ways. Our understanding of where and how to "attach" annotations of various types to base data, and how to let others search (query), view, or track annotations across applications, researchers, time, and migrations remains sketchy. A near complete scoping report provides an exploratory literature review of previous research concerning databases and annotation and defines a range of research topics that the DCC might pursue in this area. Among the problems worthy of investigating is the extent to which forms of annotation can be predicted when metadata formats or databases are designed.

Archiving and appraisal

Since digital curation is an active and costly process the long term preservation of most digital objects is unlikely to happen. Objects which are to be preserved will need to be selected for investment and care. Traditional archiving methods include some form of appraisal to determine what data to preserve and when to preserve it; these methods must also be in place for digital data. Current research also concerns techniques to appraise and archive efficiently large databases that undergo regular change, such as genome or protein databases that grow rapidly as biological research moves forward.

Provenance and data quality

The value of digital curation is lessened if the evidence as to the origin and integrity of the data is lost or unknown. Current DCC work in this area includes developing formal models to support the description of the state of the problem and to extend our understanding of the problems that arise when data are copied from one database to another or when tracking where the data has come from that has not been adequately documented (its provenance not suitably clear). As these formal models emerge they will assist those developing software tools and standards for tracking, exchanging and managing the provenance of data as it is transferred between databases.

Metadata extraction

Current digital preservation processes require extensive human intervention in selection, validation, description, assigning unique identifiers, data management, migration, and delivery of desired content. The investment of human labour currently required does not scale to the number or complexity of the digital content that needs to be curated. A case in point is preservation metadata, which is an essential part of the information infrastructure necessary to support all the processes in digital preservation. To make the preservation of digital objects more viable, automatic or semi-automated creation and authoring of the metadata must be addressed. The challenge is the development of methods and approaches for creation of metadata supporting the understandability of digital objects.

Legal issues

Legislation and the variations in legal frameworks across national boundaries pose obstacles to digital curation, but also create opportunities. DCC led investigations include examination of the impact of licensing, intellectual property rights, issues surrounding digital rights in databases, and the implications for how courts view digital information and the ways these views can be shaped by how digital materials are curated.

Networks of trusted repositories

With the diversity in repository implementations currently available the issues involved in their coordination and integration and applicability to particular resource classes require investigation. Attempts have been made to identify the necessary attributes of trusted repositories, and this has promoted the recognition that research is needed to address how multiple repositories might co-operate within a global curation network. Co-operation offers many potential benefits. For example, successful networks may enable repositories to co-operate on the selection of content and on the development of technical approaches to curation, thereby helping to reduce costs and duplication of effort. They may also be able to help reduce the risks of institutional impermanence by supporting the distribution and replication of data across multiple institutions.

Economic cost-benefit analysis

Curation activities appear to be expensive. This view must be balanced, however, with the economic and social costs of losing digital assets, some of which may be unique and impossible to recreate. Research in this area will use economic analysis and modelling techniques to explore the cost-benefits of digital curation.

Performance and optimisation

The data avalanche hitting many scientific disciplines imposes new responsibilities on data creators, users and curators. When data volumes are too large for researchers to download data sets to analyse on their desktops, the data centre is not only the place where the data are stored but where they must be analysed. Data curators must therefore provide the environment within which users can upload and run analysis code safely, while ensuring that such analysis does not have an impact on the integrity of the data collection itself. For an increasing number of studies this analysis of code will require access to data located elsewhere. If this is to be possible the data centre must be capable of storing intermediate data for users. This is an area where data curation research is likely to interface directly with Grid computing.

Services

The digital information creating, using, and curating communities require a range of practical services if they are to have any chance of securing, using, and enabling the longevity of digital information. The provision of effective, practical services for the stakeholder community is a high priority for the DCC. Work in each of the other task areas reflects this drive to deliver tools and services which will enable communities to curate their digital resources effectively. The project has created a conduit from research and development into the establishment and delivery of world-class services and products.

The DCC is *not* itself a digital repository. The intention is to provide services and guidance that helps data centres and repositories be more productive in the selection, acquisition, ingest and exploitation of data in their care by current and future users, both with existing computing and the developing e-science/Grid infrastructure. Initial tasks are directed towards intelligence gathering across our communities about needs, practices and provision, and emergent standards. In subsequent work the intention is to provide information on tools for automating the ingest and curation process, and on mechanisms that support continuing use of digital resources, data interchange and support preservation.

Help desk and FAQs

Edinburgh's experience through EDINA provided the consortium with evidence of the merits of a single point of contact for queries, staffed throughout office hours and skilled at logging, tracking, filtering and passing queries onto the appropriate expert staff. As the DCC develops such information resources as Frequently Asked Questions (FAQs) and its programme of site visits many queries will either be answered through our website or by the help desk staff pointing callers to the FAQ in which the answer may be found. The latter is greatly facilitated by the development of FAQs. The help desk also provides support for online bookings for training events and site visits.

Advisory Service

Prior to the DCC's conception there was no single central source of expertise on digital curation and preservation matters serving the United Kingdom's further education (FE) and higher education (HE) communities from research to learning and teaching to administration. Scientists, researchers, and librarians need to search across many sources of information if they wish to identify best practice, find the latest standards information, or locate a recommendation on appropriate file formats. One of the major benefits of the Advisory Service in partnership with the help desk and Web portal, is that it delivers a one-stop-shop where relevant information is gathered, evaluated, and presented to the community. Consortium partners and their DCC staff are responsible for contributing to this activity to ensure that the wide range of expertise based

at each partner site is mobilised for the maximum benefit of the community. The service is co-ordinated from HATII at the University of Glasgow, which has gained through ERPANET [18] extensive experience in leading advisory services in digital preservation. The DCC's professional advisors offer advice on strategic, technical, practical and legal aspects of data curation and preservation. The service provides authoritative answers to digital curation and preservation questions, and makes accessible examples of best practice. The advisory service works closely with every part of the DCC, responding to queries that cannot be dealt with by the help desk or extant FAQs. The advisory service tracks DCC (and other) research and development activities in order to keep up-to-date with innovative research, it monitors standards and technology watch outputs, ensures that the community is informed about leading edge activities, and repackages R&D materials appropriately for practitioners. Using site visits and focus groups the advisory service engages with the wider community in order to gather user requirements and feedback on services which will inform work on set-up and gap analysis, tailors services to the different sectoral/disciplinary groups, develops consensus on best practice, disseminates information to the widest possible audience (e.g. Current Awareness Bulletin), and works towards the formation of a curation community that pro-actively shares expertise. More in-depth or tailored support is available where required, through a range of additional services such as institutional site visits, so that the DCC can accommodate the varied requirements of its users in a cost effective and efficient manner.

Standards

The DCC operates a standard watch covering relevant existing and emerging standards, identified through interaction with users and researchers. The DCC staff actively contribute to the definition of emerging standards by working with the community and standards establishing bodies (e.g. ISO), organising associates groups around new standard developments, and initiating standardisation definition groups where gaps have been identified.

Audit and certification

Institutions are increasingly recognising the need to provide their user communities with access to trusted repositories. Irrespective of whether they construct and manage these themselves or rely on outsourced services institutions need mechanisms to validate the trusted status of repositories. The Online Computer Library Center (OCLC) [28] and the Research Libraries Group (RLG) [29] in their *Attributes of a Trusted Digital Repository* [30] paper, have proposed a high level model for the design, delivery, and maintenance of a digital repository, and RLG and NARA [33] are progressing towards certification requirements for establishing and selecting reliable digital information repositories. The DCC partnership will continue collaboration with RLG, (CCLRC is already represented on this working group), to ensure that in the development of audit and certification standards international consensus is favoured. The DCC will use its testbeds to support the development of these audit and certification standards. Although there is growing awareness of the certifiable characteristics (e.g. activities, attributes, functions, processes) of repositories the mechanisms for audit and the process by which certificates would be issued (and revoked) remain to be agreed. The Centre offers guidance on self-audit and self-certification and will conduct independent audits and issue certificates to repositories which meet its guidelines.

Developmental workshops and professional development

Digital curation is an immature discipline. There are some areas in which our knowledge is secure enough to provide training but in others the community is still developing its expertise. The DCC has planned a series of workshops and training events to reflect existing knowledge and practices, and to enable the community to work together to build new understanding in other areas.

Our programme of training workshops and seminars on best practice, strategy and applications is designed to promote participation across the diverse range of curation and preservation need communities in FE and HE institutions. This provides a network of learning for demand-driven training and continuing professional development (CPD) related to preservation and curation. This work has begun by identifying where and what further practitioner training and staff development is required. Identification and discussion of key digital curation issues highlighted by Research and Development will follow, with subsequent collaboration fostered between private industry and public organisations. Our initial workshops cover Persistent Identifiers [12], Digital Repositories [11], and Costing Models [8] for Digital Curation.

Digital Curation Manual

The DCC is committed to the publication of a range of resources to further assist institutions, data centres and repositories with their digital curation efforts. At the forefront of these is the *Digital Curation Manual* [10], a world-class resource. The Digital Curation

Manual is being constructed from fascicules written by international experts overseen by leading researchers and practitioners in the area of digital curation. Among the forty-five initial topics are Appraisal and Selection, Costs, Freedom of Information, Interoperability, the OAIS Model, Preservation Strategies, and Open Source. A full list of areas that the curation manual aims to cover can be found at the DCC web site. To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of some fifty fascicules, but as new topics emerge and older topics require more detailed coverage more might be added.

The DCC recognises that the manual is very much focused on the needs of technical staff. For many topics, a less in-depth insight is being offered by the DCC briefing papers, which have been designed to meet the needs of senior managers. These supply quick and high level overviews of the topics that are explored in technical detail in the Curation Manual.

Development

A central, and demanding, task is to transform research products into services for those who are responsible for digital preservation and data curation activities within institutions. This requires both attention to the definition of desirable services and a concerted programme of development activities. There is need here to look to, and gain leverage from extant research product as much as from research that is now in progress, and to build on work successfully carried out by others as well as by staff within the DCC. In line with concern for the longevity of information content as well as the longevity of digital media, early priority within the development programme has been given to 'representation information', as defined within the OAIS Reference Model. The DCC anticipates that the results of research suitable for transformation into products or services will be some years away. Therefore, in the first instance the development team is looking outside the core DCC for research results which can be delivered as products, methods, or services. These development activities are focussed on making available tools and services which support various aspects of the OAIS Reference Model, and is, wherever possible, standards based in its approach.

Representation Information Registry and Repository Architecture

Long term access to Representation Information is essential for the curation and preservation of digital information as it documents the structure and semantics of the ways in which digital data are stored and provides a method for accessing the content of digital objects. Representation Information is used here in the context of the OAIS model, meaning any and all information that may be required to access the information content of a data object. Representation Information is not necessarily the original or official software access method or format specification, but can take the form of anything that allows the information content of a digital object to be interpreted. The current focus of development within this area lies with the design and implementation of a distributed system of Representation Information Registry/Repositories (DCC-RR). As automation of process is essential if digital curation is to prove cost effective efforts to facilitate interoperability and automated processing of representation information lies at the core of this work.

In order to implement the DCC-RR, the open-source freebXML [20] registry software is being extended to meet the functional and architectural needs. Client software is under development to provide a variety of methods to interface with the repository contents. Descriptive and preservation metadata schemas are under investigation, focussed on enabling the contextual identification of objects and maintaining a record of their authenticity and provenance. To facilitate interoperability, standardised packaging structures are under development, these will support distributed and federated registry/repositories. Although the current focus of development lies with the technical infrastructure, there remain many structural, organisational, and rights issues associated with building such a repository that must be addressed. These will enable the selection of Representation Information for inclusion in the registry, profiling of usage constraints, mechanisms to ensure the contents are maintained and up-to-date, and adhere to internationally agreed standards.

Project Integration and Community Involvement

To broaden the scope of development and the comprehensiveness of the services provided, the development team intend to build upon the resources and findings of existing projects. For example, interoperability is essential for Digital Curation, with value identified in making DCC services as easy to integrate as possible into existing projects which use or service digital information. Such interoperability will aid the development of a global network of Representation Information repositories. Investigations are being undertaken into how Representation Information may fit into the architecture of digital repositories in order to provide such a global,

distributed infrastructure.

The dynamic nature of the contents of the DCC-RR means that maintaining and updating the information resource will require continuous investment of effort and resource. The development team will look at ways to foster and encourage community interaction and contribution in order to improve the implementation, the functionality provided, and the information contained within.

Tool development

In order to best exploit the DCC-RR's potential services that will sit on top of the software implementation are needed. These services will be primarily software driven, taking the form of external tools and extension modules providing functionality that builds upon the core registry utilities. The DCC-RR requires that a label is associated with every digital object, which acts as a pointer towards relevant Representation Information stored within the repository. Tools must be selected or developed that identifies both the requirements of a digital object and the Representation Information that satisfies these requirements. A label containing persistent identifiers pointing into the DCC-RR should be produced from this resultant set of information. Additional utilities will provide functionality to browse, search, view, and update contents. Tools to provide data description are in active development, and the development team plan to investigate methods to identify the format of unknown files, verify whether a file is the format it purports to be, assess the viability and success of transformation, and identify the risk that particular representations may pose for a special digital object.

Community Development and Population

The dynamic nature of the contents of the DCC-RR means that maintaining and updating the information collection will require significant work. Community interaction is welcomed and actively encouraged to assist in the improvement of the implementation and the functionality it provides.

Standards

The development work is based on international standards such as OAIS and others in the OAIS roadmap, and there are ongoing efforts to influence the development of several other international standards. One developing standard of particular interest deals with OAIS certification.

Testbeds

A number of factors influence the suitability of a curation software solution for a particular purpose. These factors include the cost, the effectiveness of technical solutions such as migration or emulation, techniques for metadata creation and management, and verification of the authenticity of digital objects. A test bed framework is under development which will operate on a variety of hardware and software configurations. The test beds are expected to enable the assessment of testing of these tools and processes to a high standard of QA to enable their distribution and usage by designated user communities. Wherever possible, this work will be carried out in collaboration with the original tool developers.

Outreach & Associates Network

At the forefront of all the DCC's interactions with its stakeholder communities is our Outreach team, who through a variety of activities are committed to the promotion of the DCC's message, its services and its research and development successes. The Centre must achieve high visibility if it is to have a measurable impact on the communities of practice and of policy that it aims to serve. Like NeSC, the DCC is a community forum that serves as an environment for informed discussion and debate on key issues, and aims to facilitate the sharing of expertise and skills. Outreach and dissemination work provides the DCC with the opportunity to listen and by capturing details about individual user and community needs, to reflect them in our research, development, and services activities. Close contact is anticipated with other key organisations including the Digital Preservation Coalition (DPC) [16] and the DELOS Network of Excellence [13,14] in order to co-ordinate timings and to run joint events.

User requirements

The DCC has also commissioned a study to investigate user requirements for digital curation as part of the Outreach activity. The study has used a mix of methodologies such as focus groups and interviews with key individuals who are working with data sets, researchers drawn from various disciplines, and policy and funder representatives. The outcomes from the study include an outline taxonomy of curation users and organisations and an overview of requirements, which will inform the planning of services and research and development activities of the DCC. The results suggest that there is a clear need for further information and expertise in a wide range of areas, including professional development and training, technical advice and guidance on best practice, research on topics such as annotation, economic cost-benefits analyses and pragmatic case studies.

Web portal

This 'virtual point of presence' is the main access route for users to discover information about the DCC and UK-led work in this area, including best practice guides, and digital curation resources. The DCC Web site showcases the wide variety of user communities whose work and challenges the DCC addresses. The Web portal represents views of different disciplines, with language about digital curation and preservation tailored to their needs and circumstances. The portal is the first to establish a cross disciplinary set of tools and advice that encompasses the full range of scholarly enquiry. Reciprocal links to and from related sites are being established, thus leveraging the ways we can reach out to our target user communities. The Web team is working closely with the Advisory Service, and this collaboration is yielding online documentation including best practice guides and other deliverables, that show the expertise gained through interaction with those in the field and provide the community with access to preservation information.

International Journal of Digital Curation

There is currently no high-quality peer reviewed journal in digital curation. If digital curation is to have a significant intellectual focus it must have a journal. As a result the production and delivery of an electronic journal dedicated to digital curation and preservation research, service development issues is seen as a key element of the outreach activity [23]. This e-journal will not be a dissemination and promotion channel for the DCC and its services. It will instead aim to provide a publishing focal point for researchers to present outstanding contributions to the field; of course, we hope our independent peer reviewers will consider the outputs of the DCC research team suitable for publication in the journal, but it will be their call.

Associates Network

The DCC Associates Network [9] makes the DCC partnership more pervasive; it brings together prominent members from UK data creating and managing organisations, leading data curators overseas, supranational standards agencies, and representatives of sectors of UK industry and commerce involved in digital curation. Members for the Associates Network continue to be sought. The benefits of membership include: financial savings for DCC events, early access to outputs from leading-edge research and development, support from Advisory Services, tutorials, training courses, input into user needs definition and service design, access to a range of collaborative tools, including a web-based forum, and the ability to post news and events information on the DCC's web site.

Way Forward

Our vision of a successful DCC is one that is coherent and international in its outlook which promotes world-class research and ensures that the fruits of that research will serve the purposes of scholarship and science in the UK. By combining acknowledged research strengths with proven service organisations, we seek to achieve the virtuous circle (see figure 2 above). By engaging varied communities of practice through our network of associates, and by gaining leverage from their expertise through mechanisms such as secondment, fellowships and part-time assignment, we ensure the relevance of our research, and inform the advice we can offer our users. This lays the grounds for: leadership and advocacy, continuing professional development, and promoting digital curation and preservation.

The consortium partners recognise that the viability of the DCC depends upon it gathering the right level of expertise, making that expertise available to the widest community, and demonstrating long-term commitment to the provision of research, development, services, and outreach. The DCC business plan sets incremental stage targets towards achieving long-term sustainability and it accommodates review of the DCC's progress towards achieving these targets. Encompassing vision, expertise and an acute awareness of the essential role of effective curation in all our digital activities, the DCC aims to be the embodiment of its vision for digital curation, and to succeed in its goals to be a standard-bearer for best practice in an area that is relevant to every individual, institution and organisation that relies upon and uses digital information.

Acknowledgements

The authors wish to thank all their colleagues in the DCC and on this occasion to acknowledge particularly Andrew McHugh (HATII, University of Glasgow), Robin Rice (Edinburgh University Data Library), and Adam Rusbridge (HATII, University of Glasgow) for their contributions. Many individuals played a crucial role in the design of the Digital Curation Centre and are active in helping it to deliver on its mission, among these are Charlotte Waelde (Co-Director of the AHRC Research Centre for Studies in Intellectual Property and Technology Law), and Bob Mann (University of Edinburgh and the National eScience Centre). The authors wish to thank Professor Tony Hey (Director of

the UK e-Science Core Programme of the EPSRC), and Neil Beagrie (JISC), for their support and work to construct the funding platform that has made the DCC possible.

References

[1] Access Grid, http://www.accessgrid.org/ [Accessed: 10 May 2005, 16:44]

[2] AHRC Research Centre for Studies in Intellectual Property and Technology Law, http://www.law.ed.ac.uk/ahrb/index.asp [Accessed: 9 May 2005, 11:40]

[3] Ariadne Magazine, http://www.ariadne.ac.uk/ [Accessed: 9 May 2005, 12:00]

[4] Biotechnology and Biological Sciences Research Council, http://www.bbsrc.ac.uk/ [Accessed: 9 May 2005, 12:14]

[5] *CCSDS,* 2002, *Reference Model for an Open Archival Information System (OAIS)*, http://www.ccsds.org/documents/650x0b1.pdf

[6] Council for the Central Laboratory of Research Councils, http://www.cclrc.ac.uk [Accessed: 9 May 2005, 11:29]

[7] Database Group at Edinburgh, http://www.lfcs.inf.ed.ac.uk/research/database/dbs.html [Accessed: 9 May 2005, 11:39]

[8] DCC and DPC Joint Workshop on Digital Curation Cost Models, http://www.dcc.ac.uk/cmworkshop.html [Accessed: 11 May 2005, 13:31]

[9] DCC Associates Network, http://www.dcc.ac.uk/network.html [Accessed: 9 May 2005, 12:26]

[10] DCC Digital Curation Manual, http://www.dcc.ac.uk/

[11] DCC Workshop on Long-Term Curation within Digital Repositories Workshop, http://www.dcc.ac.uk/drworkshop.html [Accessed: 11 May 2005, 13:30]

[12] DCC Workshop Workshop on Persistent Identifiers, http://www.dcc.ac.uk/piworkshop.html [Accessed: 11 May 2005, 13:30]

[13] Delos Network of Excellence on Digital Libraries, http://www.delos.info [Accessed: 9 May 2005

[14] DELOS Network of Excellence on Digital Libraries, Digital Preservation Cluster, http://www.dpc.delos.info [Accessed: 10 May 2005, 9:25]

[15] Digital Curation Centre, http://www.dcc.ac.uk [Accessed: 9 May 2005, 11:20]

[16] Digital Preservation Coalition, http://www.dpconline.org [Accessed: 11 May 2005

[17] EDINA, http://edina.ac.uk [Accessed: 9 May 2005, 11:47]

[18] Electronic Resource Preservation and Access Network (ERPANET), http://www.erpanet.org [Accessed: 9 May 2005, 11:53]

[19] Engineering and Physical Sciences Research Council, http://www.epsrc.ac.uk/ [Accessed: 9 May 2005, 11:25]

[20] Freebxml, http://www.freebxml.org [Accessed: 11 May 2005, 16:40]

[21] Giaretta, D, et.al. 2005, *Draft DCC Approach to Digital Curation*, http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCApproachToCuration?rev=1.1.9

[22] Humanities Advanced Technology and Information Institute (HATII), http://www.hatii.arts.gla.ac.uk [Accessed: 9 May 2005, 11:25]

[23] *International Journal of Digital Curation*, http://www.ijdc.net [Accessed: 9 May 2005, 12:05]

[24] Joint Information Systems Committee (JISC), http://www.jisc.ac.uk [Accessed: 9 May 2005]

[25] JISC Circular 06/03 (Revised) Digital Curation Centre http://www.jisc.ac.uk/index.cfm?name=funding_digcentre [Accessed: 9 May 2005, 12:24]

[26] JISC Press Release, 2004, New UK Centre to Communicate Across Time, http://www.jisc.ac.uk/index.cfm?name=pr_dcc_051104

[27] National e-Science Centre, http://www.nesc.ac.uk [Accessed: 9 May 2005, 11:29]

[28] Online Computer Library Center, http://www.oclc.org/ [Accessed: 9 May 2005, 12:28]

[29] Research Libraries Group, http://www.rlg.org [Accessed: 9 May 2005, 12:29]

[30] RLG-OCLC, 2002, *Attributes of a Trusted Repository*, http://www.rlg.org/longterm/repositories.pdf [Accessed: 9 May 2005, 12:30]

[31] School of Informatics at the University of Edinburgh, http://www.inf.ed.ac.uk [Accessed: 9 May 2005, 11:30]

[32] UKOLN, http://www.ukoln.ac.uk [Accessed: 9 May 2005, 11:29]

[33] US National Archives and Records Administration,http://www.archives.gov [Accessed: 9 May 2005, 12:31]

[34] University of Bath, http://www.bath.ac.uk [Accessed: 9 May 2005, 11:29]

[35] University of Edinburgh, http://www.ed.ac.uk [Accessed: 9 May 2005, 11:25]

[36] University of Glasgow, http://www.gla.ac.uk [Accessed: 9 May 2005, 11:30]