



**Report of the Warwick Workshop - 7 & 8 November 2005**

## **Digital Curation and Preservation: Defining the research agenda for the next decade**

### **1 Introduction**

Over recent years it has become clear that accessing and preserving digital data is increasingly important across a wide range of scientific, artistic and cultural activities. There has been a growing recognition of the need to address the fragility and accessibility of the digital information collected in all aspects of our lives. Access to digital information lies at the heart of the scientific and technical innovation vital for modern economies. Supporting Society's growing dependence on digital information for its smooth operation provides a real urgency for this task and a wide range of initiatives are already underway at both the national and international level to tackle the many aspects of the end-to-end digital preservation "lifecycle".

A European Task Force has been established to address the issue of maintaining permanent access to the digital Scientific Record. Its aim is to draw up a strategic action programme that addresses the full range of technical, organisational, economic, legal and social issues including an infrastructure to support permanent access. The proposed outline research programme<sup>1</sup> is to be put forward for consideration as part of the 7<sup>th</sup> Framework Programme for RTD<sup>2</sup> of the EU.

The aim of the European Research Infrastructures<sup>3</sup> conference (December 2005) is to facilitate the circulation and maintenance of information that is relevant to policy makers and other stakeholders in view of FP7 in the domain of research infrastructures. The conference provides a key milestone for the co-ordinated approach to research infrastructures in Europe as developed by the European Strategy Forum on Research Infrastructures (ESFRI)<sup>4</sup> and should contribute to clarifying the long-term scientific needs in relation to European research infrastructures. This conference should help to clarify how a Research Infrastructure can generate impacts not only on scientific issues but also on socio-economic and policy development. It is clear, more than ever before, that there is a need to overcome the fragmentation of policies in Europe.

The UK Government's "Science and Innovation Investment Framework: 2004 - 2014" identifies systematic preservation of digital information as an important component of the information infrastructure<sup>5</sup>. "Curation and preservation"<sup>6</sup> of digital information is now being taken forward by the Office of Science and Technology (OST) working group on e-Infrastructure<sup>7</sup> as one of six key components of a future national e-

---

<sup>1</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/RD\\_proposal\\_def.pdf](http://www.dcc.ac.uk/training/warwick_2005/RD_proposal_def.pdf)

<sup>2</sup> [http://www.cordis.lu/en/src/f\\_009\\_en.htm](http://www.cordis.lu/en/src/f_009_en.htm)

<sup>3</sup> <http://www.nottingham.ac.uk/ecriuk/>

<sup>4</sup> <http://www.cordis.lu/esfri/home.html>

<sup>5</sup> [http://www.hm-treasury.gov.uk/media/33A/AB/spend04\\_sciencedoc\\_1\\_090704.pdf](http://www.hm-treasury.gov.uk/media/33A/AB/spend04_sciencedoc_1_090704.pdf)

<sup>6</sup> Note that the term "digital curation" normally includes "preservation".

<sup>7</sup> <http://www.e-irg.org/>

Infrastructure. The OST working group is charged with mapping out relevant developments, gaps and challenges in digital curation and preservation over the next 10 years.

In the light of these activities it was timely to bring together, at this Warwick workshop, national and international experts across the full spectrum of the digital lifecycle to assist this process by mapping out the current state of play and future agenda and provide valuable input to the working group. A particular focus of this workshop was the research agenda for digital preservation and curation; what were the major challenges and gaps in current developments and what more was needed to tackle them? This national and international context was presented by Dr Malcolm Read, Executive Secretary of the Joint Information Systems Committee (JISC), Professor John Wood<sup>8</sup>, the Chief Executive of the Council for the Central Laboratory of the Research Councils (CCLRC) and chair of ESFRI, and Peter Tindemans<sup>9</sup>, who chairs the Task Force on Permanent Access from a European perspective, in their opening addresses to the workshop. Neil Beagrie, BL/JISC Partnership Manager gave the concluding address. He outlined the recommendations of the previous Warwick workshop held in 1999<sup>10</sup> and reviewed the progress that had been made in implementing them over the subsequent five years.

This Warwick 2005 workshop<sup>11</sup> will complement the Task Force on Permanent Access to the Records of Science's research agenda, extending considerations explicitly into the "sociological" and policy areas, while taking an independent view of the technical programme, thereby extending and clarifying a number of research issues.

The workshop focussed on three main strands in parallel breakout sessions and group discussions. Each breakout group considered one of the following topics: **Curation Services and Technologies**<sup>12</sup>, **Drivers and Barriers** (policy issues); and **Data Life Cycle Management** (process issues). Each session was chaired by a leading expert on the topic and the groups were asked to consider the topic in relation to the following categories: the scope and definition of each topic, the current state of play nationally and internationally; what the vision was likely to be over the next 5 to 10 years; what we needed to do to achieve this vision; what were the dependencies on which achievement of the vision would be based, and what were the priorities. The breakout groups are described in more detail in the Appendices; each group tackled and presented their work in the way which was felt to be most appropriate.

The results of all the groups are summarised in section 2. These are organised, for ease of reading, into three areas: (1) those topics which recurred in the groups, and which therefore deserve special note, (2) those which were specific to individual groups and (3) a number concerned with general policy and Infrastructure development.

A more detailed report from each group is provided in Appendices 1-3, and all the presentations are available on the web site.

---

<sup>8</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/JVWWarwickpres20051107.ppt](http://www.dcc.ac.uk/training/warwick_2005/JVWWarwickpres20051107.ppt)

<sup>9</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/TindemansWarwick07-11-05.ppt](http://www.dcc.ac.uk/training/warwick_2005/TindemansWarwick07-11-05.ppt)

<sup>10</sup> <http://www.leeds.ac.uk/cedars/OTHER/warwick2.htm>

<sup>11</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/](http://www.dcc.ac.uk/training/warwick_2005/)

<sup>12</sup> also referred to as "Distributed Architectures" for historical reasons

## 2 Infrastructure Development and Research Themes

### 2.1 Common research issues identified across all three discussion groups

|  |  |
|--|--|
| <b>Discovery and location</b>                                  | 1. Adopt or develop an agreed, <b>persistent, actionable, identifier for digital objects</b> , with associated name resolvers which are themselves persistent. |
|  | 2. Continue to develop <b>search and discovery tools</b> in partnership with relevant user groups.   |
|  | 3. Develop more <b>detailed Data Models</b> for each domain and abstract out intra-domain and inter-domain commonalities.                                      |
| <b>Trust</b>   | 4. Develop and integrate <b>DRM, provenance and authenticity</b> checking into ingest processes.   |
|  | 5. Prototype and test national certification “badges” as <b>prototypes of certification processes</b> .  |
| <b>Cost</b>  | 6. Continuing <b>data collection and modelling of costs</b> , with adequately complex parameterisation, over the life-cycle of different data types.           |
| <b>Automation and Virtualisation</b>                           | 7. Develop language to <b>describe data policy</b> demands and processes, together with associated support systems.  |
|  | 8. Develop <b>collection oriented description</b> and transfer techniques.   |
|  | 9. Develop data description tools and associated generic migration applications to <b>facilitate automation</b> .  |
|  | 10. Develop <b>standardised intermediate forms</b> with sets of coder/decoder pairs to and from specific common formats.                                       |
|  | 11. Develop <b>code generation tools</b> for automatically creating software for format migration.   |
|  | 12. Develop techniques to allow <b>data virtualisation of common science objects</b> , with at least some discipline specific extensions.                      |
|  | 13. <b>Management and policy specifications</b> will be need to be formalised and virtualised.   |
|  | 14. Further <b>virtualisation of knowledge</b> – including developments of interoperable and maintainable ontologies.  |
| 15. Develop <b>automatic processes for metadata extraction</b> |  |

## 2.2 Specific research topics

|                |   |
|----------------|---|
| Virtualisation | 16. Continuing work on ways of <b>describing information all the way from the bits upwards</b> , in standardised ways – “virtualisation”. Work is needed on each of the identified layers in section A1.2.              |
|                | 17. <b>Knowledge virtualization</b> involving Ontologies and other Semantic Web developments are required to enable the characterization of the applicability of a set of relationships across a set of semantic terms. |
|                | 18. Develop use of <b>data format description languages</b> to characterize the structures present within a digital record, independently of the original creation application.   |
|                | 19. It is important to make significant progress on dealing with <b>dynamic data including databases</b> , and object behaviour.  |
|                | 20. <b>Representation Information tools</b> , probably via layers of virtualisation to allow appropriate normalisation, including mature tools for dealing with dynamic data including databases.                       |
|                | 21. Additional work on <b>preservation strategies and support tools</b> , from emulation to virtualisation.   |
|                | 22. Develop increasingly powerful virtualisation <b>tools</b> and techniques, with a particular emphasis on knowledge technologies.   |
| Automation     | 23. Develop protocols and information management exchange mechanisms, including synchronisation techniques for indices etc., to <b>support federations</b> .  |
|                | 24. <b>Standardised APIs</b> for applications and data integration techniques   |
|                | 25. Fuller development of <b>workflow systems and process definition</b> and control.   |
| Support        | 26. Develop simple semantic <b>descriptions of Designated Communities</b> .   |
|                | 27. <b>Standardise Registry/Repositories for Representation Information</b> to facilitate sharing.  |
|                | 28. Develop <b>methodologies and services for archiving personal collections</b> of digital materials.  |
| Hardware       | 29. Develop and <b>standardise interfaces</b> to allow “pluggable” storage hardware systems.  |
|                | 30. <b>Standardise archive storage API</b> i.e. standardised storage virtualisation.  |
|                | 31. Develop <b>certification processes for storage systems</b> .  |
|                | 32. Undertake research to characterise types of read and <b>transmission errors</b> and the development of techniques which detect and potentially correct them.  |

## 2.3 Policy and infrastructure development

|  |  |
|--|--|
| <b>National and international infrastructure</b> | 33. The need for a national (and international) <b>roadmap for an infrastructure to support long-term curation and preservation</b> , which is underpinned by common policies and standards, that address the roles and responsibilities of the various stakeholder groups |
|  | 34. The need for a <b>significant increase in investment</b> at national and international level   |
|  | 35. More support for <b>collaborative activities</b> , at national and international level, is needed  |
| <b>Cross-cultural and cross-disciplinary</b>     | 36. A clearer understanding of the needs of diverse disciplines and <b>encouragement</b> for cross-disciplinary programmes is required   |
|  | 37. More <b>training and accreditation</b> is required for information professionals   |
|  | 38. More <b>advocacy is needed in support of changing the research culture</b> to embrace the challenges, and invest time up front, in the curation and preservation process   |
| <b>Partnerships</b>                              | 39. Work in partnership with <b>commercial system providers and with key interested parties</b> such as CERN and others, on error levels and developing affordable scalability   |
|  | 40. Build <b>accredited community resources</b> , such as public data bases, with a cachet of contributing data.   |
|  | 41. Work with data generating community, and their funders, to <b>encourage the adoption of standards</b> .  |
| <b>Support Infrastructure</b>                    | 42. A clearer understanding of information management with respect to legal issues such as <b>IPR and trust</b> is needed  |
|  | 43. The need for <b>more best practice guidelines</b>  |
|  | 44. Support for <b>persistent resolver services</b> for persistent identifiers   |
|  | 45. Support for <b>shared registries/repositories</b> of various types of metadata, particularly Representation Information and curation tools   |
| <b>Standards</b>                                 | 46. Progress <b>Ingest standards</b> (e.g. follow-on from PAIMAS standard).  |
|  | 47. Progress <b>Audit and Certification</b> draft to full ISO standard.  |
|  | 48. Set up <b>accreditation process</b> under the supervision of an international body.  |

### **3 Conclusions**

The research topics and policy issues identified in this workshop provide a valuable addition to the discussion of research priorities over the next decade. They are, although having a more explicit emphasis on virtualisation and identifying a number of broader policy and infrastructure issues, consistent with and complementary to the European research programme<sup>13</sup> proposed by the Task Force on Permanent Access<sup>14</sup>, and the Strategic Action Programme 2006-2010<sup>15</sup>. The latter summarises the research elements as covering:

- A core set of physical digital archives
- Conditions to ensure proper archiving, interoperability and long-term preservation for long-term accessibility of data
- A Framework for Metadata (including Representation Information), a Framework for Persistent Identifiers, and a number of registries
- Cost-effective long-term preservation methods and services
- Digital Access and Rights Management
- Mechanism for developing and testing implementation tools, techniques and services
- Certification service providers
- A common European accreditation mechanism

It is pleasing to note that the current workshop has validated and extended the programme of the Task Force on Permanent Access, and has provided further detailed thinking to help advance that research agenda, and complement a number of earlier reports.<sup>16</sup>

Report Co-ordinators:

David Giaretta, Heather Weaver (CCLRC)

*The event was sponsored by the JISC, the CCLRC, the DCC and the BL.*

*The sponsors would like to express their thanks to the members of the Steering Committee for their help in structuring and organising the event and to all delegates for their enthusiastic participation at the event itself.*

---

<sup>13</sup> <http://tfpa.kb.nl/Proposal%20Research%20and%20Development.doc>

<sup>14</sup> Further information is available at the web site of the European Task Force Permanent Access at <http://tfpa.kb.nl/>

<sup>15</sup> See <http://tfpa.kb.nl/Strategic%20Action%20Programme.pdf>

<sup>16</sup> See for example *Invest to Save*, Report and Recommendations of the NSF-DELOS, Working Group on Digital Archiving and Preservation, (Hedstrom and Ross, 2003) <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf>

## **Appendix 1: Curation Services and Technologies<sup>17</sup>**

This session was chaired by David Giaretta, Associate Director for Development of the Digital Curation Centre, and the focus was the services and technologies which are needed to support curation.

Presentations by Olaf Barring<sup>18</sup> (CERN) and Steve Hughes<sup>19</sup> (Planetary Data System/JPL) provided an overview of the current state of the art, together with reasonable extrapolations to the 5 year horizon. Views to the 10 year horizon were provided by Bruce Wright<sup>20</sup> (UK Met Office) and Reagan Moore<sup>21</sup> (San Diego Supercomputer Centre). Finally Lou Reich<sup>22</sup> (NASA/CSC) provided a summary of the problems that need to be solved, with a specific OAIS Reference Model<sup>23</sup> point of view.

### **A 1.2: Where are we now?**

A broad brush review of the state of the art for scientific (and other) archives suggests the following points.

- Most archives which contain primary research data are domain focussed. There is a consensus that domain experts are best placed to provide support for the users of the archive data, and moreover are best placed to define new data products and user services. Closely related to this is the observation that such archives are, in the overwhelming number of cases, of limited lifetime (but there are notable exceptions). The lifetime is related to that of, for example, the instrument which was used to gather the data, or to a particular research project for which the data has been created or gathered.
- The limitations of individual, local, filesystems have been overcome, and (small) multiple Petabyte storage systems, centralised or distributed, consisting of 10's of millions of files, are just starting to come into common use, although access patterns and efficient data mining tools for storage systems of this size are not yet available. Error rates and error correction codes for commercially available hardware are insufficiently strong for these volumes of data.
- The regular, and costly, upgrades of repository hardware and of software are generally still very painful, especially migration of information holdings from old system to new. The same considerations apply for the movement of holdings between organisations.
- Outsourcing of storage is beginning to happen, and funders are beginning to require at least a limited time of guaranteed access to data produced by research that they fund. The need for some kind of certification has been enunciated for more than a decade but only now are the first steps being taken.
- Interoperability in terms of catalogue harvesting and search tools, while neither widespread nor well integrated, is becoming increasingly common. Deeper

---

<sup>17</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/sessions/distributed\\_architectures](http://www.dcc.ac.uk/training/warwick_2005/sessions/distributed_architectures)

<sup>18</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/CERN-DM.ppt](http://www.dcc.ac.uk/training/warwick_2005/CERN-DM.ppt)

<sup>19</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/dcc\\_hughes.ppt](http://www.dcc.ac.uk/training/warwick_2005/dcc_hughes.ppt)

<sup>20</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DigitalCurationWorkshop-20051107.ppt](http://www.dcc.ac.uk/training/warwick_2005/DigitalCurationWorkshop-20051107.ppt)

<sup>21</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/future-warwick.ppt](http://www.dcc.ac.uk/training/warwick_2005/future-warwick.ppt)

<sup>22</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/warwick.ppt](http://www.dcc.ac.uk/training/warwick_2005/warwick.ppt)

<sup>23</sup> <http://public.ccsds.org/publications/archive/650x0b1.pdf>

interoperability in terms of combining data elements from a variety of data sets, is limited to small groups, and is largely left to client tools, which tend to have idiosyncratic user interfaces and highly prescribed operations.

- Federation between archives, in the sense of being viewable as a single system, is coming closer, but currently it takes the form of “bolt-on” systems rather than formal agreements and integrated systems.
- Authentication virtualization mechanisms are available for managing user identity independently of the administrative domains which are utilized by a preservation environment, but use tends to be limited within domains or projects.
- Workflow virtualization mechanisms are now appearing that rely on grid technology to manage interactions with remote computing systems, but is not yet fully developed. Where format migrations are performed, either for storage within the archive or on-demand conversions for users, it remains something which is non-standard.
- The ingestion of data/information into an archive is normally a hand-crafted process for each dataset.
- Preservation methodologies based on virtualization are now appearing; this involves the extraction of digital records from their creation environment and their import into a preservation environment. In the process, virtualization mechanisms are used to remove all dependencies on the original creation environment.
- Explicit statements and monitoring of Designated Communities are rare.
- Storage virtualization is now available through Storage Area Networks. Versions of this technology are now appearing that will work over wide area networks.
- Archive developers are starting to look beyond the virtualisation of storage, but it is recognised that to build a “future-proof” system is going to be very hard. XML has been discussed as providing a self-describing, future-proof, format but this is now widely recognised as unrealistic for anything other than the simplest data.
- Data virtualization mechanisms are appearing in the form of data grid technology. Data grids enable the management of shared collection properties independently of the distributed storage systems in which the data reside.
- A number of digital object identifiers are in use, such as DOI and various types of URIs, and each has its committed advocates. None can provide absolute guarantees of long-term resolvability.

### **A 1.3: 5 years time**

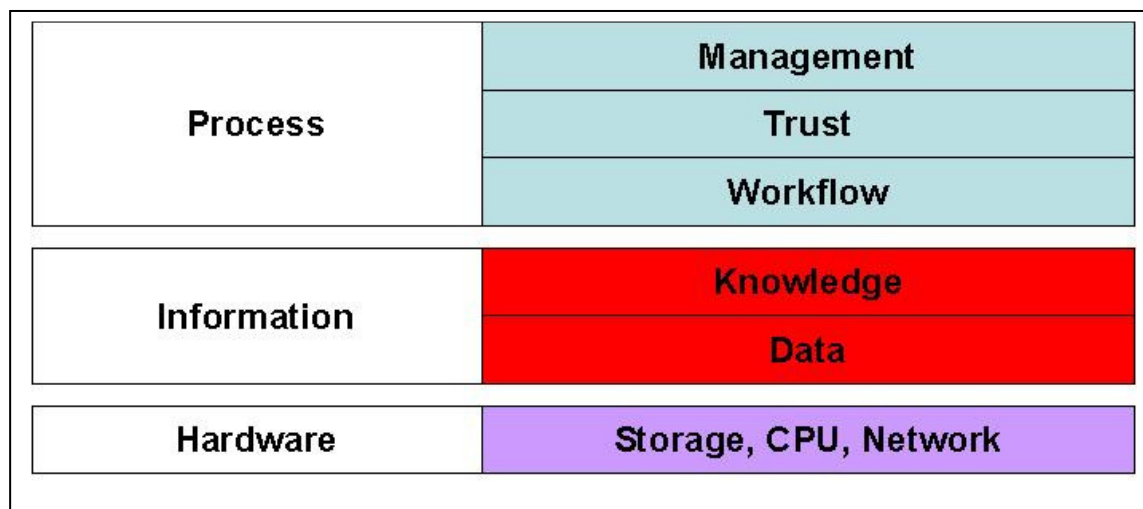
A conservative extrapolation to the next 5 years suggests that many of the limitations described in the previous section could reasonably be expected to be lifted. It is expected that progress will be made most effectively in the areas outlined below through an incremental, rather than a “big bang”, approach.

Underpinning each of the activities should be cost/benefit analyses.

The maintainability and preservability of the preservation information (Representation Information etc), systems and services themselves should also be a primary concern.

Note that, for clarity, there is a limited amount of duplication between sections under the “How to get there”.

Virtualisation is an underlying theme, with a layering model illustrated as follows:



| <b>Where we would like to be</b>  | <b>How to get there</b>   |
|---|---|
| Massively scalable architectures, with centralised services and distributed storage of many 10's of Petabytes, with several billion files, should be fairly common, while error levels for commercial systems should be acceptable. | Work in partnership with commercial system providers and with key interested parties such as CERN and others, on error levels and developing affordable scalability.<br><br>Undertake research to characterise types of read and transmission errors and the development of techniques which detect and potentially correct them. |
| Data should be identifiable independently of storage solution.  | Adopt or develop an agreed, persistent, actionable, identifier for digital object, with associated name resolvers which are themselves persistent.  |
| Archives will probably still be domain focussed but with a wider and more distributed search and discovery mechanism.   | Continue to develop search and discovery tools in partnership with relevant user groups.<br><br>Develop more detailed Data Models for each domain and abstract out intra-domain and inter-domain commonalities.   |
| Improved automation should also be available for ingest and validation.   | Develop tools for automatically capturing/creating Representation Information and PDI from data holdings.<br><br>Work with data generating community, and their funders, to encourage the adoption of standards.<br><br>Develop and integrate DRM, provenance   |

|   |  |
|---|--|
|   | and authenticity checking into ingest processes.   |
| Process automation should also have progressed on automated application of high level statement of policies on activities such as media refreshment.  | Develop language to describe data policy demands and processes, together with associated support systems.  |
| In OAIS terms we should be able to better define the knowledge base of a Designated Community in order to automate the demand for adequate Representation Information.  | Develop simple semantic descriptions of Designated Communities.  |
| <p>The set of virtualization layers needed for management of distributed data should be recognized and widely accepted. Layers for virtualisation include: Management, Trust, Workflow and Systems, Knowledge, Data, Hardware.</p> <p>Trust virtualization enables the management of authenticity and authorization across administrative domains.</p> <p>Policy virtualization enables the characterization of the preservation policies independently of the implementation choice.</p> <p>Data virtualization enables the continued management of preservation environment properties while the underlying technology evolves.</p> <p>Tools for producing Representation Information should be available.</p> <p>Service Level Agreement driven services e.g. storage systems, should be available.</p> <p>Federations of large numbers of repositories should become common.</p> <p>Infrastructure for sharing the effort of preservation of information should be commonly used.</p> | <p>Continuing work on ways of describing information all the way from the bits upwards, in standardised ways – “virtualisation”. Work is needed on each of the identified layers.</p> <p>Some specific points about Representation Information are:</p> <ul style="list-style-type: none"> <li>• Knowledge virtualization involving Ontologies and other Semantic Web developments are required to enable the characterization of the applicability of a set of relationships across a set of semantic terms.</li> <li>• It is important to make significant progress on dealing with dynamic data including databases, and object behaviour.</li> <li>• Develop use of data format description languages to characterize the structures present within a digital record, independently of the original creation application.</li> </ul> <p>Standardise Registry/Repositories for Representation Information to facilitate sharing.</p> <p>Additional work on preservation strategies and support tools, from emulation to virtualisation.</p> |
| <p>Easy movement of objects or collections between repositories.</p> <p>Much greater automation should also be available for processing digital information.</p>  | <p>Develop collection oriented description and transfer techniques.</p> <p>There are several potential strands for this:</p> <ul style="list-style-type: none"> <li>• Develop data description tools and</li> </ul>  |

|  |   |
|--|---|
|  | <p>associated generic migration applications to facilitate automation.</p> <ul style="list-style-type: none"> <li>• Develop canonical intermediate forms with sets of coder/decoder pairs to and from specific common formats.</li> <li>• Develop code generation tools for automatically creating software for format migration.</li> </ul> <p>Develop techniques to allow data virtualisation of common science objects, with at least some discipline specific extensions.</p> |
| <p>An accreditation and certification system should be available, based on an international standard and backed at international and governmental level.</p> | <p>Prototype and test national certification “badges” as prototypes of certification processes.</p> <p>Progress Audit and Certification draft to full ISO standard.</p> <p>Set up accreditation process under the supervision of an international body.</p>   |

**A 1.4: 10 years time**

In the period up to the 10 year horizon one would expect an increase in the level of automation possible, based on Knowledge technologies, increased capacity and increased sharing and interoperability.

| <b>Where we would like to be</b>   | <b>How to get there</b>  |
|--|--|
| <p>Massively scalable storage systems covering the range 100’s Petabytes to Exabytes, with essentially unlimited numbers of files, should be available without heroic efforts, nor should heroic and painful efforts be demanded to cope with changes in underlying hardware and technologies. Costs of both hardware and software systems should also be affordable, and s/w systems should be robust, seamless and stable.</p> | <p>Further work with commercial systems providers and key service providers and user groups.</p> <p>Develop and standardise interfaces to allow “pluggable” storage hardware systems.</p> <p>Standardise archive storage API i.e. standardised storage virtualisation.</p> <p>Certification processes for storage systems.</p> |
| <p>“Deep” cross-disciplinary interoperability should be seamless, even at an international level, and scalable federations of archives and also federations of federations of archives should be common, while query, search and discovery should be able to be specified in natural language.</p>   | <p>Develop increasingly powerful virtualisation tools and techniques, with a particular emphasis on knowledge technologies.</p> <p>Develop protocols and information management exchange mechanisms, including synchronisation techniques for indices etc., to support federations.</p>  |

|  |   |
|--|---|
|  | Management and policy specifications will be need to be formalised and virtualised.   |
| Knowledge and management virtualisation should be mature and should support sophisticated information integration with client tools as well as via archive services. | Continued virtualisation of knowledge – including developments of interoperable and maintainable ontologies.<br>Standardised APIs for applications and data integration techniques.<br>Fuller development of workflow systems and process definition and control. |
| Full support of Representation Information and the linkage to the Designated Community's knowledge base should also be mature.                                       | Yet more Representation Information tools, probably via layers of virtualisation to allow appropriate normalisation.<br>Must include mature tools for dealing with dynamic data including databases.  |
| Accurate cost predictive estimates of preservation activities looking 10 - 20 years ahead.   | Continuing data collection and modelling of cost data.<br>Cost/benefit modelling with complex parameterisations.  |
| Preservable and evolvable preservation systems are available   | Further develop virtualisation model (including ontology) evolution, plus dynamic models and tools for classification of new instances.   |

### **A 1.5: Dependencies**

- Cost/benefit analyses need to tie in with technology changes
- Organisational and policy developments need to be formalised and encoded
- Growth rate of storage, CPU, bandwidth assumed to hold (but there is a worry that it might not)
  - Storage
    - Depends on commercial opportunities. Bit-error rates demanded commercially may not be adequate for Exabyte scale data
    - New technologies such as holographic storage may change the market dramatically<sup>24</sup>
  - Need to maintain balance between CPU and I/O demands

More effective access to archives must go hand in hand with preservation requirements and available resource constraints

---

<sup>24</sup> See [http://www.manifest-tech.com/media\\_dvd/dvd\\_holo.htm](http://www.manifest-tech.com/media_dvd/dvd_holo.htm) - 300GB holographic disks are promised for late 2006, with capacities of more than 1TB following shortly.

## **Appendix 2: Drivers and Barriers Session**

The Drivers and Barriers session<sup>25,26</sup> was chaired by Lynne Brindley, Chief Executive of the British Library, on day one and Chris Rusbridge, Director of the Digital Curation Centre, on day two. The session started with a presentation by Professor Christine Borgman<sup>27</sup> of UCLA on the value chain of scholarly information, which was underpinned by networks of publications, data, records and composite objects. Professor Laurie Hunter<sup>28</sup> then looked at digital information as an intangible asset and discussed its structure and characteristics in comparison with other assets. Finally Robert Sharpe<sup>29</sup> of Tessella, a commercial organisation undertaking a UK Digital Preservation Needs Assessment on behalf of the Digital Preservation Coalition, gave an overview of the “sticks” and “carrots” behind organisational efforts at digital preservation.

The discussion that followed identified seven broad areas where delegates felt significant effort was needed to ensure long-term curation of digital information that would underpin a wide range of scientific, artistic and cultural activities:

- Incentives for researcher engagement
- Understanding of disciplinary differences
- Business cases
- Access and rights framework for intellectual property to support preservation over the full lifecycle of digital information
- Good and bad case studies
- Modernisation of the metadata (lifecycle and tools)
- Workforce capacity building

These themes were then combined and merged into four main areas which were addressed in four smaller groups.

### ***A 2.2: Incentives for researcher engagement and workforce capacity building***

#### **Where are we now?**

There is in general a low level of awareness of the need for digital preservation and a lack of researchers' participation. Many of them often feel confused about roles and responsibilities, as well as about copyright and intellectual property issues. Data citations are recognised in very few areas and the reward systems for researchers do not match the incentives for digital preservation. Moreover there is no explicit link between incentives and work force capacity.

#### **Where do we want to be in 10 years time?**

---

<sup>25</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/sessions/drivers\\_and\\_barriers](http://www.dcc.ac.uk/training/warwick_2005/sessions/drivers_and_barriers)

<sup>26</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DriversAndBarriersSummary.ppt](http://www.dcc.ac.uk/training/warwick_2005/DriversAndBarriersSummary.ppt)

<sup>27</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/BorgmanDriversBarriers.ppt](http://www.dcc.ac.uk/training/warwick_2005/BorgmanDriversBarriers.ppt)

<sup>28</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/LHunterDriversBarriers.ppt](http://www.dcc.ac.uk/training/warwick_2005/LHunterDriversBarriers.ppt)

<sup>29</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/TessellaDriversBarriers.ppt](http://www.dcc.ac.uk/training/warwick_2005/TessellaDriversBarriers.ppt)

A widespread appreciation of digital preservation would provide an environment in which long term re-use of digital data becomes a key activity of academic life, with clearly defined roles and responsibilities for digital preservation, and where researchers are equipped with the necessary skills. As a result, citation of data becomes mainstream alongside citation of literature, leading to much more data-led research and new types of science. Data becomes capital in the research enterprise.

### **How do we get there (the research questions)?**

- We need good examples, such as exemplar projects, to capture the hearts and minds of individual researchers and to show incentives to organisations, so that collaborations can be formed to tackle the problems jointly.
- Build accredited community resources, such as public data bases, with a cachet of contributing data.
- Map out generic processes in digital preservation and identify those which are discipline specific. A two-pronged approach is needed, on the one hand improving understanding of the incentives for participation and on the other testing new reward systems (RAE, hiring etc).
- Increase capacity by training existing and new workforces through innovative undergraduate and Masters programmes which focus on data management, to establish a cadre of preservation professionals who understand data; at the same time provide generic research training to existing workforce in digital preservation.

## ***A 2.3: Disciplinary differences***

### **Where are we now?**

There is currently little understanding in any detail of the real heterogeneity of data management and preservation practices across different disciplines. Differences can be seen in *deposit* arrangements and requirements; *preservation* arrangements; in the use of data and information; observational versus experimental data, and external sources; organisational and institutional; research process and methods; and different levels of workforce skills.

### **Where do we want to be in 10 years time?**

A much improved understanding of the real requirements of different disciplines will lead to a cultural change in the attitude towards data sharing, which will lead to fruitful interactions within and between various disciplines and sub-disciplines. In addition, a developed and interoperable infrastructure will be in place, nationally and internationally, which focuses on re-access and re-use of data. There will also be a common framework for policies and procedures, with clearly defined roles and responsibilities.

### **How do we get there (the research questions)?**

- The bed-rock of research in this area is to understand in more detail the sociology of preserving and sharing information. This will include understanding better disciplinary differences, and in particular those requirements that are fundamental versus those that are primarily historical. For a cultural change to take place, it is important to involve key stakeholders and resource providers and for them to drive this process.

- To encourage inter and intra-disciplinary interactions, issue-driven research programmes (cross-disciplinary?) need to be funded.
- To build the desired infrastructure, repository development is essential, so are national and international partnerships.
- It is also necessary to distinguish separate preservation and usage (access) layers within the infrastructure and build the respective services that fit the specific purposes.
- A road map is needed at national and institutional level to define the roles and responsibilities of national organisations, funding bodies and institutions.
- We need to start working on a common framework for policies and procedures, and develop and train the workforce so that they possess the adequate and necessary skills.

## ***A 2.4: Business cases, rights and responsibilities***

### **Where are we now?**

The situation now is that there is an increasing understanding of the nature and general dimensions of costs involved in digital preservation. However most of this is related to ingest, rather than the different stages of the full lifecycle. Risk management is a well-researched area and a lot can be learnt and applied when managing the risk related to digital preservation. The value of digital data and that of the activity of digital preservation (which provides future value to a current digital object that would otherwise be lost) is, however, not as well explored.

### **Where do we want to be in 10 years time?**

What we would like to see in 10 years time is in the first instance a much clearer understanding of IPR issues, as well as roles and responsibilities. Furthermore, any legal change that has impact on digital preservation will be well understood and advocated accordingly. Good practice guidance, applicable to different sectors, will be widely available and there will be a real preservation structure available, deployed and used; research will be driving the cost bases downwards.

Most of all, however, it will be possible to build convincing business cases for digital preservation; business cases that will allow Board-level investment decisions to be made on a rational basis. In research, the nature of “Public Good” and its relationship with digital preservation will be clear.

### **How do we get there (the research questions)?**

- We need to build a business case and research the legal aspects - case study based.
- Identify rich case studies on cost, risk and benefits of digital preservation with wide stakeholder engagement, covering a range of scenarios.
- Further modelling based on those case studies to extend to new areas and opportunities.
- Apply practical R&D in preservation which is applicable to different communities.

- Take a four-pronged approach that engages the funders, with top-down effort at the management and policy level, bottom-up participation from the researchers, and a practical deployment perspective.

## **A 2.5: Modernisation of the metadata**

### **Where are we now?**

Metadata have many aspects, but in this context two levels are important: discovery level<sup>30</sup> and domain-specific content level<sup>31</sup>. The current metadata standards mainly deal with the discovery level, providing ways to describe content, allowing automated discovery, however it does not allow automatic processing of the data itself. Domain specific content metadata should not only be machine-readable, but may be automatically machine-processable. Domain ontologies are examples of this more detailed content description, however, that area of work is not yet mature.

There is currently some staged metadata collection (creation, ingest, update, use etc.) but there is little clarity on responsibility and/or authority and there is little integration of these patchily collected metadata.

We do not know how to estimate the cost of producing adequate metadata. We do know, however, that the increasing demands for ever more metadata mean that traditional hand-crafted approaches (as in Library MARC records, for instance) are no longer affordable.

### **Where do we want to be in 10 years time?**

It is likely that metadata standards will be in place at all levels, more widely deployed and implemented much more scalably than now. Some key metadata practices will have changed beyond all recognition. Metadata standards at discovery level will be containers that allow descriptions of content and context (semantic web/RDF/logic) and they will be both machine “understandable” and automatically processable. Again, metadata standards at a domain-specific content level will provide deep syntax and semantics, including domain ontologies and will be populated and linked. Staged metadata collection will be supported with easy-to-use interfaces and automated metadata creation, capture and update will be widely available. Metadata will be captured closer to the point of resource creation, as part of the creation process (at a time when the required information is cheaply accessible). More metadata will be inferred automatically from the characteristics of the resource. There will be clearly defined responsibility and authority with all metadata collected at different stages integrated and cross-referenced, and a good understanding of the value of metadata for curation and of the value of particular fields of metadata and ways to quantify these values.

### **How do we get there (the research questions)?**

- Develop machine-“understandable” discovery metadata – container, content and context, supported by general domain ontologies.
- Develop domain-specific metadata – machine-understandable container, content and context, supported by specific (but linked) domain ontologies.

---

<sup>30</sup> in OAI terms this would be Descriptive Information used by Finding Aids

<sup>31</sup> in OAI terms this would include Representation Information and Preservation Description Information

- Define the process of metadata collection; make available metadata information and develop IT support systems.
- Develop and utilise automated metadata collection tools.
- Create systems for metadata capture at resource creation.
- Develop responsibility / authority models and implement them.
- Apply integration methodologies and develop models for value estimation and value proving.

## **Appendix 3: Data Lifecycle Management**

This Breakout Group<sup>32,33</sup> was chaired by Dr Anne Trefethen<sup>34</sup>, the Director of the e-Science Core Programme. It focussed on an e-researcher's perspective on digital curation and preservation throughout the scholarly knowledge lifecycle, informed by practical issues from funders and collaborative services. Presentations were given on the Scholarly Lifecycle by Dr Jeremy Frey, from the University of Southampton; Mark Thorley<sup>35</sup> on behalf of RCUK, who looked at the data sharing and curation policies across the UK Research Councils<sup>36</sup>; and Life Cycle Collection Management and costing by Helen Shenton<sup>37</sup> of the British Library.

### **A 3.2: Where are we now?**

We are generating material faster than we are taking care of it, without thought for the long term value. The pace of technology means faster lock out of material. There is a high expectation by users in the networked world that they can use search tools like Google to locate information. Researchers are now becoming concerned about data management and are beginning to realise the value and need for personal archiving, reinvention and replication. There is a lack of tools and education – for both professionals and researchers - coupled with a lack of review mechanisms for scientific and other digital archives. Academic literacy is changing and there is a growing democratisation of the publication process. More requirements will be made of data from scholarly publishing.

### **A 3.3: Where do we want to be in 10 years time?**

Ten years will bring much larger scale interoperation of data resources, easily discovered and seamlessly used – across data types – across the lifecycle of data – across silos of data – in the context of the broader scholarly knowledge cycle. Automatic tools for semantic information import and export, autonomic curation (e.g. agents) and provenance capture will be deployed. All types of multimedia will be as easily indexed and searched as text today. We anticipate provision of larger and faster, trusted and secure, storage and high bandwidth network allowing rapid search.

### **A 3.4: How do we get there?**

- We need to develop an understanding of future publishing, in terms of culture, mechanisms, and rewards.
- Have a unified, clearly understood, policy framework.
- Develop and utilise tools and technologies to support the lifecycle, including appraisal mechanisms for selection, ingest, metadata creation, curation etc.
- Establish trusted, ethical and legal frameworks within lifecycle management

---

<sup>32</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/sessions/data\\_lifecycle\\_management](http://www.dcc.ac.uk/training/warwick_2005/sessions/data_lifecycle_management)

<sup>33</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DCPWarwick\\_DLM\\_summary81105.ppt](http://www.dcc.ac.uk/training/warwick_2005/DCPWarwick_DLM_summary81105.ppt)

<sup>34</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DCP\\_DLM\\_Breakout\\_session.ppt](http://www.dcc.ac.uk/training/warwick_2005/DCP_DLM_Breakout_session.ppt)

<sup>35</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DCPWarwick\\_DLM\\_MarkThorley\\_71105.ppt](http://www.dcc.ac.uk/training/warwick_2005/DCPWarwick_DLM_MarkThorley_71105.ppt)

<sup>36</sup> The CCLRC, as one of the sponsors of the event, is using the outputs from the workshop to inform the development of a digital curation policy for the research data from created from its facilities

<sup>37</sup> [http://www.dcc.ac.uk/training/warwick\\_2005/DCPWarwick\\_DLM\\_HelenShenton\\_71105.ppt](http://www.dcc.ac.uk/training/warwick_2005/DCPWarwick_DLM_HelenShenton_71105.ppt)

- Develop economic models for data management
- Establish “Knowledge” curators who can inform and support the research culture.
- Co-operation and collaboration mechanisms to provide support across stakeholder groups and provide long- term investment in data centres and research infrastructures.

### ***A 3.5: Dependencies***

Long-term investment is required, along with policies which outline roles and responsibilities.

Enabling technology is required along with the people to develop it and educate the users. This training and education requires a blend of expertise. We need to encourage stakeholder collaboration, both nationally and internationally, and address the legal and ethical requirements

### ***A 3.6: What are the priorities?***

- Provide stability and guidance to researchers, via prototypes.
- Encourage co-operation which can lead to interoperability.
- Establish a standards framework with certification and trust mechanisms.
- Create motivation models and build automation for scale.
- We need a research programme to develop required technologies and to understand the social and cultural aspects.
- Develop life-cycle cost and economic models for digital data.

## Appendix 4: Attendees

| <b>First name</b> | <b>Surname</b> | <b>Organisation</b>   | <b>Group</b> | <b>Role</b>    |
|-------------------|----------------|-----------------------|--------------|----------------|
| Matthew           | Addis          | U Southampton         |              |                |
| Alison            | Allden         | University of Bristol | DLM          |                |
| Kevin             | Ashley         | ULCC                  | CST          |                |
| Steve             | Bailey         | JISC Records Manager  | DLM          |                |
| Olof              | Barring        | CERN                  | CST          |                |
| Neil              | Beagrie        | The British Library   | DLM          | Org, SC        |
| Juan              | Bicarregui     | CCLRC                 | DLM          | SC             |
| Christine         | Borgman        | UCLA                  | DB           |                |
| Richard           | Boulderstone   | British Library       | CST          |                |
| Michael           | Bright         | ESRC                  | CST          |                |
| Lynne             | Brindley       | British Library       | DB           | Chair          |
| Peter             | Brophy         | U Manchester          |              |                |
| Adrian            | Brown          | TNA                   | CST          | Org            |
| Peter             | Burnhill       | EDINA                 | DLM          |                |
| Helen             | Campbell       | AEDC                  | CST          |                |
| Julia             | Chruszcz       | MIMAS                 | DLM          |                |
| David             | Corney         | CCLRC                 | CST          | SC             |
| David             | Dawson         | MLA                   | DB           |                |
| Stuart            | Dempster       | JISC                  | DLM          |                |
| Jeremy            | Frey           | U Southampton         | DLM          |                |
| Luigi             | Fusco          | ESA/ESRIN             | DB           |                |
| Neil              | Geddes         | CCRLC                 | DB           | SC             |
| David             | Giaretta       | CCLRC                 | CST          | Chair, Org, SC |
| Jerry             | Giles          | NGDC                  | DB           |                |
| Mariella          | Guercio        | U Urbino              | CST          |                |
| Sara              | Hassen         | JISC                  |              | Reporter       |
| Jessie            | Hey            | U Southampton         | DLM          | Org, SC        |
| Helen             | Hockx-Yu       | JISC                  | DB           | Org, SC        |
| Stephen           | Hughes         | JPL                   | CST          |                |
| Laurie            | Hunter         | U of Glasgow          | DB           |                |
| Joseph            | Hutcheon       | JISC                  | DB           |                |
| Paul              | Jeffreys       | Oxford U CS           | CST          |                |
| Keith             | Jeffery        | CCLRC                 | DB           |                |
| Maggie            | Jones          | DPC                   | DLM          |                |
| Michael           | Jubb           | RLN                   | DB           |                |
| William           | Kilbride       | AHDS                  | DLM          |                |
| Kerstin           | Kleese-Van-Dam | CCLRC                 | DLM          | SC             |
| Michael           | Lautenschlager | WDC, Germany          | CST          |                |
| Martin            | Lewis          | Sheffield University  | DB           |                |
| Philip            | Lord           | DAC                   | DB           |                |
| Liz               | Lyon           | UKOLN                 |              |                |

|           |            |  |     |                |
|-----------|------------|--|-----|----------------|
| Alison    | Macdonald  | DAC  | CST |                |
| Reagan    | Moore      | SDSC   | CST |                |
| Carlos    | Oliveira   | European Commission                                | DB  |                |
| Norman    | Paskin     | IDF  | CST |                |
| Sam       | Pepler     | BADC   | CST |                |
| Philip    | Pothen     | JISC   |     | Reporter       |
| Adrian    | Pugh       | RCUK   | DLM |                |
| Malcolm   | Read       | JISC   | DB  |                |
| Lou       | Reich      | NASA/CSC   | CST |                |
| Gill      | Ross       | MET Office   | CST |                |
| Seamus    | Ross       | U of Glasgow                                       | CST |                |
| Chris     | Rusbridge  | DCC  | DB  | Chair, Org, SC |
| Kevin     | Schurer    | UKDA   | DB  |                |
| Robert    | Sharpe     | Tessella   | DB  |                |
| Helen     | Shenton    | BL   | DLM |                |
| Pauline   | Simpson    | NERC   | DLM |                |
| Mike      | Smorul     | U Maryland   | CST |                |
| Matthew   | Stiff      | CEH  |     |                |
| Allan     | Sudlow     | MRC  | DLM | SC             |
| Simon     | Tanner     | KCL DCS  | DB  |                |
| Mark      | Thorley    | NERC   | DLM |                |
| Peter     | Tindemans  | Council for the European Spallation Source Project | DLM |                |
| Anne      | Trefethen  | UK e-Science Programme                             | DLM | Chair          |
| Nick      | Trigg      | CCLRC  | CST |                |
| Colin     | Venters    | U Manchester                                       | CST |                |
| Charlotte | Waelde     | AHRB   | DB  |                |
| Martin    | Waller     | Tessella   | DB  |                |
| Heather   | Weaver     | CCLRC  | DLM | SC             |
| Murray    | Weston     | BUFVC  | DB  |                |
| Andrew    | Wilson     | AHDS   | DB  |                |
| Astrid    | Wissenburg | ESRC   | DLM | SC             |
| John      | Wood       | CCLRC  | DB  |                |
| Bruce     | Wright     | MET Office   | CST |                |

| Key:   |                                    |       |                 |
|--------|------------------------------------|-------|-----------------|
| Groups |                                    | Roles |                 |
| DB:    | Drivers and Barriers Group         | Org   | Group organiser |
| DLM    | Data Life-cycle management         | Chair | Session chair   |
| CST    | Curation Services and Technologies | SC    | Steering cttee  |