E-INFRASTRUCTURE STRATEGY FOR RESEARCH: FINAL REPORT FROM THE OSI PRESERVATION AND CURATION WORKING GROUP

NEIL BEAGRIE

JANUARY 2007

Preface

In July 2004 the Treasury, Department of Trade and Industry (DTI) and the Department for Education and Skills (DfES) published the "Science and Innovation Investment Framework 2004-014", which set out the government's ambitions for UK science and innovation over that period, in particular their contribution to economic growth and public services.

A section of the Framework addressed the need for an e-infrastructure for research. It proposed an Office for Science and Innovation (OSI) lead steering group to focus discussion and assess requirements for its development.

To implement this recommendation, the OSI steering group was formed and commissioned a study to help inform the process of achieving the objectives set out in the Framework document. The study was to tasked with establishing a high-level "road map" of the current provision of the UK's "e-Infrastructure" to support research, and in doing so help define the development this infrastructure.

The steering group subsequently formed six working groups to develop this road map of the e-Infrastructure in greater detail in specific areas. These working groups were tasked with producing the following reports:

1. Middleware and AAA (authentication, authorization, accounting) and digital rights management
2. Networks, compute power and storage hardware
3. Preservation and curation
4. Search and navigation
5. Information creation and data creation
6. Virtual research communities

The individual reports are intended to represent the informed opinion of the working groups and their contributors and to guide discussion and future development of the e-infrastructure. The working groups have worked closely together. Although each report is free-standing, a synthesis of all the reports and major issues has also been produced which will provide a framework for departmental bids in the next Comprehensive Spending Review and for future planning and development of the e-infrastructure for research.

Prue Backway
Office for Science and Innovation
Department of Trade and Industry

Contents

Executive Summary

Over the next 10 years the move to a digital knowledge economy should largely have been completed. Government, research, individuals, and businesses will be dependent on digital information. The Science and Innovation Investment Framework has argued that over the next decade the growing UK research base must have ready and efficient access to digital information of all kinds such as experimental data sets, journals, theses, conference proceedings and patents. This is the life blood of research and innovation but presents a number of major risks due to unresolved challenges in their long-term management.

This report summarises the work and recommendations of the preservation and curation sub-group formed by the DTI steering group to address the issues and challenges related to digital preservation and curation. Its focus is primarily on the public sector particularly government and academic research. Key findings and messages are:

- There will be dramatic growth in digital research data and publications over the next decade and a requirement to transform information provision so that UK researchers can benefit from the new research opportunities it will create;
- There are major challenges in the preservation and curation of digital information;
- Developing a persistent information infrastructure and new research and development programmes will be critical to delivery of the Science and Innovation Strategy and Transformational Government agendas;
- Where disciplinary data centres and services exist they represent approx 1.4-1.5% of total research expenditure (excluding indirect overheads/full economic costs);
- Digital preservation and curation are complex issues requiring a strategic approach to policy and development of the national infrastructure;
- Major investments are being made in the USA in their information and Cyber infrastructure: the relative position of the UK to the USA is weakening;
- We expect major new industries and opportunities to open up in these areas;
- This is not just a UK government problem to solve and there will need to be close partnership with industry in developing and delivering solutions.

Key recommendations are:

- **Policy Development**. We recommend reviewing government legislation, regulations, codes of practice, and policies, to require or to emphasise existing requirements for adequate long-term protection and appropriate accessibility of valuable information from science data through to administrative data, and electronic publications.
- **Digital Preservation and Curation Research and Development Programme**. We recommend research council and DTI-funded research programmes of fundamental and applied research by universities and industry to address long-term digital preservation and curation challenges.
- **Persistent National Information Infrastructure Development Programme**. We believe there will be a requirement in an initial period for a central DTI funded national information infrastructure development programme to enable the transformation and to pump-prime the development of the information infrastructure.

1. Introduction

In July 2004 the Treasury, DTI and DfES published the “Science and Innovation Investment Framework 2004-2014” (HMSO 2004), which set out the government’s ambitions for UK science and innovation over that period, in particular their contribution to economic growth and public services. A section of the strategy addressed the need for an information e-infrastructure and proposed a DTI (OSI) led steering group to focus discussion and assess funding requirements to develop this.

There are significant opportunities but also risks in developing a knowledge economy and research culture dependent on digital information and processes. The Science and Innovation Investment Framework 2004-2014 recognised that much of the information resources needed for research is now, and increasingly, in digital form. This is excellent for rapid access but presents a number of potential risks and challenges. For example, the digital information created during the last 15 years is in various formats (versions of software and storage media) that are already obsolete or risk being so in the not too distant future. Digital information is also often transient in nature, especially when published formally or informally on websites; unless it is collected and archived it will disappear.

The term digital preservation has been used in this report for the series of managed activities necessary to address these preservation challenges and to ensure continued access to digital information for as long as necessary. Alongside digital preservation the term digital curation is increasingly being used for the actions needed to add value to and maintain digital research assets over time for current and future generations of users. The concepts of digital preservation and curation are still relatively new and usage varies between sectors and disciplines but they should be seen as closely integrated and complementary terms.

This report summarises the work and recommendations of the preservation and curation sub-group formed by the DTI steering group to address the issues and challenges related to digital preservation and curation. It is divided into five main sections. The first three are devoted to summary overviews of the current position, our vision for the ideal position in 10 years time, and our recommendations and proposals on how to get there. Finally we have gathered the evidence and detailed reports on the current position and future work required into two major appendices (appendix B and C). The evidence gathered represents a snap-shot as of March 2006. It will inevitably change over time.

Membership of the Preservation and Curation Group is given in Appendix A. The working group has utilised the Roadmap Study completed by the Digital Archiving Consultancy of the current provision of the UK’s “e-Infrastructure” to support research, the report of the November 2005 Warwick Workshop on "Digital Curation and Preservation: Defining the research agenda for the next decade", and the Digital Preservation Coalition commissioned report “Mind the Gap: Assessing digital preservation needs in the UK” together with input from the individuals and organisations represented.

Neil Beagrie
British Library/Joint Information Systems Committee
Chair OSI Preservation and Curation Working Group
November 2006

2. Summary of current position

2.1 Investment in digital preservation and curation services

Current UK provision of repositories and investment in digital preservation and curation is very uneven. Across the UK as a whole there are still many significant gaps in the provision or necessary scale of long-term facilities by discipline, sector, and size of institution. These gaps have been highlighted recently by both the DTI commissioned “Survey of the UK’s current e-Infrastructure Provision for Academic Research” (Digital Archiving Consultancy 2005) and the Digital Preservation Coalition commissioned “Mind the Gap: Assessing Digital Preservation Needs in the UK” (Tessella 2006). Large repositories have or are being developed by a few national bodies such as the British Library, the National Archives, the Ordnance Survey, and some research councils such as AHRC, CCLRC, ESRC, and NERC. The commissioned surveys suggest the need to address identified gaps and also to build on and accelerate development of existing repositories and services.

2.2 Growth trends in digital research information

Current concerns over developing provision reported by respondents to these surveys, is heightened by knowledge of the ongoing trends in the output of and user demand for digital research information. These trends are so dramatic that they are putting and will continue to put substantial pressure on the existing information infrastructure and working practices. Some examples for both data and publications are given below. More detailed examples and projections of growth trends for research information and of growth in user demand for digital research information are given in Appendix B.

2.3 Growth in volume and use of research publications

Worldwide growth in published information of both serials and monographs, and a growing shift from paper to electronic publication, are already widely recognised trends amongst research libraries. Recently the British Library in launching its new three year strategy estimated that by the year 2020, 40% of UK research monographs will be available in electronic format only, while a further 50% will be produced in both print and digital form. A mere 10% of new titles will be available in print alone by 2020. This will not impact solely on the British Library: this will be a seismic shift for the Library and its partners in publishing and the information sector as a whole (British Library 2005). While this trend has the potential to bring huge benefits to all sectors of the community, it will place an even greater urgency on the need to develop an infrastructure capable of managing these digital resources.

2.4 Growth in volume and use of research data

For scientific data, Hey and Trefethen argued that experiments and instruments currently being built will dramatically escalate the current rates and volumes of scientific data creation. They point out that e-Science data generated from sensors, satellites, high-performance computer simulations, high-throughput devices, scientific images and so on will soon dwarf all of the scientific data collected in the whole history of scientific exploration (Hey and Trefethen 2004). Major data growth is occurring not just in “big” science but also at the level of individual researchers. With the growth of digital research data there are also increasing opportunities for data sharing and re-use and achieving efficiencies in research funding (Lord et al 2005). In January 2004 national governments including that of the UK signed the OECD Declaration on Access to Research Data from Public Funding highlighting these opportunities and agreeing to work towards commonly agreed principles and guidelines (OECD 2004).

2.5 Selection for preservation and curation: the role of data and information management policies

Not all of the digital information being generated will have long-term value and need to be curated and preserved for the future. However a significant and growing proportion of it does. It is essential that decisions on selection for this preservation and curation form part of an organisational process and are not made on an ad hoc basis. This requires ongoing processes for care, and selection for retention or disposal. In many cases it may not be possible to make a decision on retention and long-term value until sometime after creation. This implies that good procures are also needed in the interim period. Data and record management and information management policies are central to this and require co-ordination of policy at a high-level. This is increasingly being recognised and addressed by research funders but as yet very few have fully formed data and information policies in place. A full overview of the current position is given in Appendix B.

There is often critical linkage between policy and investment in the creation of research data and other research outputs and policy and investment in curation and preservation. For example initial metadata creation, project documentation, permissions and access arrangements are central to future re-use and access. This underlies the need for a strategic approach to these issues by funders. Particular difficulties may arise where there are inconsistencies between short-term project funding and the medium to long-term needs of science and research. Examples of such difficulties have recently been published in relation to long-lived data (National Science Board 2005).

2.6 Investment in R&D

We note current research and development funding for digital preservation and digital curation has been limited but increasing in recent years. The majority of UK funding has been from the JISC for development projects with very little research council funding available for digital preservation or curation research. There is also a danger that the range of researchers and practitioners involved has been too narrow: there is a need for a broader research base. This observation on the research base has also been made by the NSF/DELOS working group on digital preservation (Hedstrom et al 2003). The DTI commissioned “Survey of the UK’s current e-Infrastructure Provision for Academic Research” (Digital Archiving Consultancy 2005) notes the very high international standing of past UK research and development projects in digital preservation and curation but suggests that the lead that the UK had is in danger of being lost due to insufficient ongoing investment in these areas. It is also noted that embedding in institutions and services is less than optimal if there is an over-reliance on project funding for development and investment in services.

An expert workshop convened at Warwick in November 2005 on “Digital Curation and Preservation: Defining the research agenda for the next decade” has made the case for a higher level of investment in research and development of tools for automation and scaleable architectures, and to develop organisational and economic models to meet the challenges of increasing volumes of digital information and a corresponding increase in demand for digital preservation and curation services (Giaretta et al 2006).

The requirements described above were also echoed in the DPC commissioned *Mind the Gap* report. One of the 18 identified needs was that ‘There needs to be more technical tools to help organisations perform digital preservation activities such as performing format migrations, format validation and automated metadata extraction'. The report also highlighted a number of needs targeted at increasing awareness throughout an organisation and for funding to take account of the long-term value of digital resources being produced (Tessella 2006).

2.7 Industry and public sector services

Currently the terms digital preservation and curation would not be widely understood or used across industry and the public sector services although some of the underlying issues would be familiar and solutions to them of common benefit to industry, science and the public sector.

Retention of electronic records for many decades is a growing issue related to compliance with UK and international legislation for many sectors including pharmaceuticals, aerospace, petrochemical and nuclear industries, environmental science and engineering, oil and gas companies, healthcare, financials, and the legal profession. Compliance and records retention bring in the long-term issues associated with digital preservation such as obsolescence, authenticity, and access over long time periods.

For the aerospace and engineering sectors a major emerging digital preservation challenge is one of model retention. Product behaviour is increasingly being simulated before the physical product is made, and the design models are then used as a direct input to the manufacturing machines and robots. There is a practical need to retain information on the facilities and equipment used. For buildings the loss of the building CAD model will mean that the cost of projects like re-cabling will be unpredictable. Approval for facilities involving chemical, biological or nuclear hazards will be contingent on showing that they can safely be decommissioned, and consequently that any electronic models needed for decommissioning will be available in the long term. Similarly, for complex experiments, retaining the model of the experimental system may be as important as the results.

Industry and the public sector are also influenced by the growing volumes of digital information, the requirement for greater inter-operability between different data sources and data types, and the need to add value to and exploit corporate information. Issues that need to be addressed in similar ways as for digital preservation and curation include: reducing the cost and complexity of maintaining and exploiting information through the provision of more open file formats and data exchange formats; information lifecycle management policies; and more effective mass storage. There are therefore shared benefits and interests. Emerging examples of this are current moves towards defining open specification standards for Office formats by Microsoft, standard archival file formats by the digital photographic industry, or archival PDF formats by Adobe. There is still substantial work to do in these areas and potentially valuable collaborations to advance this work can take place between the public and commercial sectors. Public sector purchasing policies and guidelines may also play a role in encouraging this transformation in a similar way to the role played by The National Archive's programme to develop functional requirements for electronic records management systems, and to approve software products against those requirements.

In November 2005, the Government published its strategy, Transformational Government, for using IT to drive forward change, improvement, and efficiency in public services. The basic agenda is about customer focus, shared services, and professionalism of delivery. A key underlying theme of Transformational Government is that our government and our society are being transformed by the use of technology. Technology is no longer just a tactical addition to conventional business models and methods but a strategic tool for change, modernisation, and efficiency in its own right. One of the drivers for action is that in many parts of society and the economy, we are moving past the “tipping point” of the adoption of IT – where the electronic version is the only version, where money, decisions, and knowledge only exist online (Cabinet Office 2005). Digital preservation is one of the big issues that information producers, information users, government and industry need to address in a digitally-based society –for government IT users it needs to be very much part of the Transformational Government agenda.

Major industries such as publishing, media, and science based commercial research also have a vested interest in supporting the long-term preservation and curation of digital information addressed by others even if they do not undertake this activity directly themselves. For example the DTI commissioned report 'Publishing in the Knowledge Economy' undertaken in partnership with the publishing associations argued that “in order for the UK to protect access to important research material and to ensure that small and not-for-profit publishers are not unfairly disadvantaged, the archiving of digital research should be organised at a national level by Government.” (DTI 2003).

Private individuals are producing an increasingly large quantity of personal, digital information, some of which resides on their own computing devices and some externally. Email, financial records, documents, audio, images and video are key examples. These “personal digital collections” are of significant value to the individuals concerned and notable examples will become available to researchers through libraries, archives and other repositories. While facilities and services for personal archiving are beginning to appear, they remain limited to providing backup and file store functionality with no genuine long-term preservation capability.

Over the next decade many innovations and opportunities for new services and businesses to support the information and knowledge management needs of industry, the public sector, and of individuals will be created. The DTI is in a position to both influence and assist interaction between industry, the public sector and scientific research and catalyse further joint work on digital preservation and curation. We believe investment in these areas within science and research also will bring major benefits to industry and the public sector. However this would not be a one-way street in terms of benefits: we believe partnerships with industry will be essential in scientific research and development and that many electronic records management practices evolving elsewhere in the public sector will provide valuable input to science and research.

2.8 International comparisons

Internationally a wide range of bodies including the International Council for Science (International Council for Science 2004), the US Library of Congress (Library of Congress 2002), the National Science Foundation and DELOS (Hedstrom et al 2003), the US National Archive and Records Administration, the OECD (OECD 2004),and the US National Science Board (National Science Board 2005), have identified an urgent strategic requirement for today’s research community and research support organisations to assume responsibility for building a robust data and information infrastructure for the future. A number of plans and national strategies have been (or are being produced) and examples from a range of countries are documented in the appendices.

For example in the USA, the National Science Foundation's (NSF) Cyberinfrastructure Council has initiated a comprehensive strategic planning process to guide the agency's investments in cyberinfrastructure. Its draft Strategic Plan (2006-2010) for Data, Data

Analysis and Visualisation advocates creating a national digital data framework consisting of a range of data collections and managed organizations networked together. An Office for Cyberinfrastructure has been created with a budget of $127 million (£73 m) potentially rising to $182.42 million (£104 m) per annum thereafter. In parallel the US Congress has directed the Library of Congress to develop and execute the National Digital Information Infrastructure and Preservation Program (NDIIPP) and provided up to $100 million (£57 m) for this purpose. The Library of Congress and the National Science Foundation (NSF) have partnered to establish a Digital Archiving and Long-Term Preservation (DIGARCH) research programme. The US National Archives and Records Administration (NARA) has also awarded Lockheed Martin a $308 million (£176 m) contract to build a permanent archives system to preserve and manage electronic records created by the federal government. The US National Science and Technology Council has identified long-term preservation and the maintenance of and access to long-lived science and engineering data collections and Federal records as one of five strategic priorities for R&D by federal agencies in the 2007 Presidential budget.

The international comparisons suggest that the UK Science and Innovation Investment Strategy currently compares favourably in terms of other national science and research policies. It also suggests that the emphasis on supporting information infrastructure as a significant part of the strategy is correct and widely recognised elsewhere. However it is worth noting there are very major investments being made in implementation of the strategies for information infrastructure in the USA. There is a strong impression that the relative position of the UK to the USA may be weakening as a result. We believe the research councils need to look very closely at information infrastructure developments in the USA.

Within Europe it also appears that the European Union will make increased investments in the areas of digital curation and preservation research with highest priority being given to funding initiatives which have a proven track record of effort and experience in this area. Further investment in relevant UK research and information infrastructure would strengthen the position of UK researchers and their ability to successfully leverage EU funding in these areas.

Finally we would note that large-scale science is increasingly an international activity in all areas. Data curation and preservation also need to be addressed collaboratively at international level.

Note: further detailed information compiled on current position is given in Appendix B.

3. Summary of ‘theoretical ideal’ situation (10 year time horizon)

Note: further detailed information compiled on ideal situation in 10 years is given in Appendix C. This is based on extracts from the Warwick Workshop final report (Giaretta et al 2006).

3.1 A vision of the future

Over the next 10 years the move to a digital knowledge economy will largely have been completed. Government, research, individuals, and businesses will be dependent on digital information. Most human knowledge will be in digital form through a combination of continuous exponential growth in “born digital” information and increased investment in digitisation of legacy information. There will be a smooth passage of knowledge from creation to repositories. Data creators can store information in appropriate formats and locations. The information will be made available where and when appropriate to other users in a form that can be readily consumed with the technology of the day and extra value can be added in an audited and reproducible fashion.

A UK information infrastructure supported by the public and private sector will have been developed incorporating a network of long-term repositories. It will have accelerated scientific progress by providing easier access, curating and preserving core information for re-use, and applying sophisticated data mining and analysis tools to reveal new knowledge. It will have positioned UK researchers and curators to adapt to and benefit from the vast increases in digital information by providing cost-effective tools for automation of processes and resource discovery.

Major new industries and innovative business services will have been stimulated in the UK by these investments and will have emerged to support the long-term digital information and knowledge management needs of industry, the public sector, and of individuals.

3.2 Culture change

Culture change will have been recognised as one of the major challenges and significant effort invested in changing perceptions ensuring there were strong incentives for researchers and other stakeholders to support transformations in working practice. A more widespread appreciation of digital preservation would provide an environment in which long term re-use of digital information 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 and responsible stewardship is embedded within the workflow of organisations creating digital resources.

Organisational, social and economic change will have been supported. Economies of scale will have been achieved through greater collaboration, shared capacity, interoperability of repositories, and by developing a national information infrastructure.

3.3 Network of repositories and policies

Long-term threats to preservation and curation of digital information arising from organisational and administrative disruption, funding instability, or lack of clarity surrounding handover of curatorial responsibility will have been addressed. This will have been achieved through development of a network of repositories and services, replication and collaboration between them, longer-term funding frameworks, and definition of different types of repository, roles, and responsibilities over the lifecycle of research information.

We will have a complex network of trusted digital repositories and policies in place across sectors and disciplines. Good practice guidance, applicable to different sectors, will be widely available and there will be a real preservation structure, deployed and used; research and new procedures will be driving the cost base downwards.

3.4 Standards, tools and services

In partnership with the private sector, investment will have introduced widely adopted open standards for file formats, metadata description, and tools and shared services for automatically monitoring digital files, identifying risks, and addressing digital preservation challenges. There will be well-developed certification and audit processes for these services and tools. Significant collaboration and partnership between research, industry, and the public sector in the UK will have underpinned knowledge transfer and the creation of innovative new business services. Automation and the development of new generic tools and processes will accommodate higher data throughput and allow greater staff productivity. Ideally over the next 10 years we will have seen an increase in the level of automation possible, based on more R&D in knowledge and preservation technologies.

3.5 Data sharing and citation

Citation of data will have become mainstream alongside citation of literature, leading to much more data-led research and new types of science. The academic reward system will provide appropriate credit and recognition for data contributions. A much improved understanding of the real requirements of different disciplines will lead to a cultural change in the attitude towards data sharing, licensing and automating access rights, 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 access and re-use of data.

Much larger scale interoperation of data resources will be available, easily discovered and seamlessly used – across data types – across the lifecycle of data – across silos of data – and 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 more easily indexed and searched than today.

3.6 Intellectual property rights (IPR)

IPR will be respected and also will enable innovation and appropriate re-use. There will be a balance between the individual, commercial, and public interests in relevant legislation and recognition of the public interest in enabling digital preservation. IPR tools will stimulate ease of access and expressions of rights and permissions and will not be a barrier to preservation activities. We will have developed science commons licences to enable data sharing with appropriate safeguards for reserved rights and citation for data creators. We will see a much clearer understanding of IPR issues. Furthermore, any legal change that has impact on digital preservation will be well understood and advocated accordingly.

3.7 Business and Costing Models

We will have a much better understanding of possible economic models, the benefits and costings for preservation and curation over the different stages of the information lifecycle for different types of digital material. We will have invested in culture change (see 3.2 above) and identified and provided appropriate incentives to support management of research outputs over their entire lifecycle. We will have a well-formed understanding of the risks and developed risk management and audit procedures for digital information.

3.8 Metadata

Metadata standards will be in place at all levels, more widely deployed and implemented, and much more scaleable 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”, automatically processable, and scaleable over very large datasets.

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). Mechanisms for simplifying metadata creation and capture will have been developed to assist this process. 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 preservation and curation and of the value of particular fields of metadata and ways to quantify these values.

3.9 International collaboration

The UK information infrastructure and participating organisations will have developed closer links with related international initiatives. The UK will work closely with European partners on appropriate infrastructure and policies within the European Union, and with a wide range of other countries notably the USA. Open standards and many tools and services will be developed in international forums and in collaboration with UK and international partners from the public and private sectors.

4. The way in which research would progress with these additional capabilities

The Science and Innovation Investment Framework argued that over the next decade the growing UK research base must have ready and efficient access to information of all kinds – such as experimental data sets, journals, theses, conference proceedings and patents. This is the life blood of research and innovation. Further investment in digital preservation and curation capabilities will significantly enhance this process and research by:

(a) Development and provision of historic information and time series;
(b) Retaining the growing published record of science including data and published articles;
(c) Promoting the re-use and long-term cost-effectiveness of resources generated by current research;
(d) Allowing the validation of research results;
(e) Technology and knowledge transfer to industry, government and other sectors;
(f) Supporting the development of a digital knowledge economy and the transfer from paper to digital processes and formats;
(g) Identifying and disseminating best practice;
(h) Providing clear advice on selection of material suitable for long-term digital preservation;
(i) Developing tools and services to aid all of the above.

Some significant steps forward in digital preservation have been made in the UK with relatively modest investment. However, over the next 10 years the pace of change, the scale and complexity of digital resources to be curated and preserved, and the user demand for them, are expected to escalate far beyond current levels. This will require a corresponding increase in investment in R&D and infrastructure to ensure the means to manage those resources most efficiently and effectively is not compromised.

Plans for future provision by key providers are unlikely to be sufficient over the period to 2014 to accommodate this scale of change and it also seems unlikely that public (or private) investment for individual institutions could expand at anything like the same rate as information growth.

Although this transfer to an increasingly digital information environment over the next decade will impose many strains and challenges it also presents some opportunities for changing procedures and delivery of services that could address these challenges.

5. Recommended proposals

We suggest significant further effort needs to be put into developing a persistent collaborative information infrastructure for digital materials and into developing the digital curation skills of researchers and information professionals in the period 2004-2014. Proposals to achieve this are identified by the working group. Without these efforts, we believe current investment in research, digitisation and digital content will not secure lasting benefits.

5.1 Policy development

Description: Amend, clarify or reinforce government legislation, regulations, codes of practice and policies to require, or emphasise existing requirements for, adequate long term protection and appropriate accessibility of valuable information of all kinds from the science record to the public and private record. The necessary culture change implications of this will be large and are covered in [4.3] below. This proposal includes finalising regulations for legal deposit of electronic publications, as well as introducing consistency of policy and practice across various research sectors e.g. agriculture, biotechnology, health and medical research. It should also include changes to ensure IPR and related legislation appropriately balance the public and private interests, and the interests of the creator/owner and the archive/preservation service. We note a number of these issues are already under consideration in terms of digital rights management by the House of Lords Science and Technology Committee.

Benefits:

- Re-use of the outcomes of major public investment;
- Retention of the nation's scientific and cultural heritage for its future research (and business) value, accountability, evidential basis of decision making;

Risks in proposal to be managed:

- Cultural change needed in organisations to achieve compliance should not be under-estimated.

Risks in doing nothing and not proceeding with the proposal:

- Failure to deliver on Freedom of Information, Legal Deposit, e-Government and other related initiatives;
- Failure to capitalise on previous public investments;
- Future expenditure to re-create lost data;
- Loss of historic and time-based data which can not be recreated.

5.2 Digital preservation and curation r&d programme

Description: research council and DTI funded research programmes of fundamental and applied research by universities and industry to address long-term digital preservation and curation challenges. Potential areas of work to be covered by this programme have been identified in the report of the Warwick workshop (see appendix C).

Benefits:

- accelerated development of business and cost models, and technologies for curation and preservation to underpin new digital processes and working methods;
- UK wealth creation by expanding and creating new markets;
- positioning of UK research, government, and industry to harness cost benefits in information handling and knowledge extraction from long-term data and information resources;

- improves UK competitive advantage in key sectors such as publishing, aerospace, pharmaceuticals, and information storage and management;
- enabling transformational change in public services; strengthening the potential for cross-sectoral collaboration and shared services;
- building on and extending early UK reputation established for world-class R&D in this area within the UK education, government, and industry sectors;
- positioning of UK researchers to leverage further funding from the European Union and private sector.

Risks in proposal to be managed:

- insufficient follow through from research to development and service implementations;
- failure to develop a sufficiently broad research base;
- insufficient engagement of industry, practitioners and end-users with research;

Risks in doing nothing and not proceeding with the proposal:

- failure to invest sufficiently in R&D in this area and missing out on the growth of new industries;
- loss of UK research and business competitiveness in key areas;
- government objectives for transformational government, e-research, and the knowledge economy will be seriously compromised.

5.3 Persistent national information infrastructure development programme

Description: Digital curation and preservation present significant cultural and organisational challenges because the needs and benefits often extend beyond a single organisation or sector and extend over long periods of time. It is impossible to achieve the broader transformation, synergies or collaborations extending beyond individual institutions or departments without a strategic approach. The overall process of delivering the information infrastructure will also be a complex process over a number of years involving articulating the vision, building capability and capacity through pilot projects and training and skills development, through to development and to sustainable service and infrastructure. For these reasons we believe there will be a requirement in an initial period between 2007-2014 for a central DTI funded national information infrastructure development programme to enable this transformation and pump-prime development of the information infrastructure needed to deliver the Science and Innovation Strategy. We envisage this national information infrastructure will develop to include:

- One or more very large-scale (petabytes) national repositories for research data;
- National discipline based research data centres and services;
- Major national digital libraries and university digital libraries and a UK legal deposit libraries electronic network;
- Federated institutional repositories based in universities and colleges;
- A UK web-archive;
- Major national digital archives and shared services for government and the public sector;
- National or regional shared repositories to meet the needs of small and medium size organisations;
- Data and publication repositories maintained by publishers and major private sector industries;
- Links to international facilities;
- Support organisations enabling advocacy, training, collaboration, knowledge transfer, and development of shared tools, registries, and other technical services.

Benefits:

- development of an information infrastructure which can support world-class research and business development in the UK;
- supporting UK wealth creation by expanding and creating new markets;
- positioning of UK research, government, and industry to harness cost benefits in information handling and knowledge extraction from long-term data and information resources;
- underpinning UK competitive advantage in key sectors such as publishing, aerospace, pharmaceuticals, and information storage and management;
- enhancing automation of processes via online services and tools;
- enabling transformational change in the national information infrastructure and public services and evolution of a UK network of trusted repositories;
- strengthening the potential for cross-sectoral collaboration;
- sharing of best practice, tools, and services in digital preservation and curation;
- developing more digital repositories across relevant sectors and organisations including national or regional provision of services to meet the needs of small and medium size organisations;
- satisfying the growing requirements across government and research for "intermediate" (5-75 years) preservation and curation or permanent retention and use as part of the UK knowledge base and record of science.

Risks in proposal to be managed:

- the requirement for appropriate management and direction of the programme;
- transitioning from capital funding to ongoing service;
- governance issues and mechanisms for shared infrastructure and services.

Risks in doing nothing and not proceeding with the proposal:

- failure to deliver on the science and innovation strategy;
- loss of UK competitiveness in research and innovation;
- inability to cope with future research information and data growth and to capitalise on the new research opportunities they will enable;
- undermining of key UK industry sectors such as publishing that need solutions to long-term digital preservation and curation;
- undermining of the UK public sector and the transformational government agenda which will require a persistent information infrastructure for long-term sustainability of e-government.

Finally in addition to the above proposals, the working group recommended that proposals for training and skills, outreach and technology transfer, inter-operability, and quality assurance should be developed by the OSI Steering Group on a shared basis across all six OSI sub-groups.

References

British Library, 2005, British Library predicts ‘switch to digital by 2020’. Press release 29th June 2005, retrieved 17 November 2006 from:
http://www.bl.uk/news/2005/pressrelease20050629.html

Cabinet Office, 2005, *Transformational Government: Enabled by Technology* (Her Majesty’s Stationary Office, London). Retrieved 17 November 2006 from:
http://www.cio.gov.uk/documents/pdf/transgov/transgov-strategy.pdf

Digital Archiving Consultancy, 2005, *Survey of the UK’s current e-Infrastructure Provision for Academic Research* (Department of Trade and Industry unpublished)

DTI, 2003, *Publishing in the Knowledge Economy*.(Department of Trade and Industry, London). Retrieved 17 November 2006 from:
http://www.dti.gov.uk/files/file13777.pdf

Giaretta, David et al, 2006, *Digital Curation and Preservation: Defining the research agenda for the next decade*. Report of the Warwick Workshop - 7 & 8 November 2005. Retrieved 17 November 2006 from:
http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf

Hedstrom, Margaret et al, 2003, *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*. Retrieved 17 November 2006 from:
http://eprints.erpanet.org/94/

Hey, Tony and Trefethen, Anne, 2003, “The Data Deluge: an e-science Perspective” in: Berman, Fran (Ed.) et al, 2003, Grid Computing: Making the Global Infrastructure a Reality, (John Wiley and Sons). Retrieved 17 November 2006 from:
http://www.ecs.soton.ac.uk/~ajgh/DataDeluge(final).pdf

HMSO 2004, *Science and innovation investment framework 2004-2014*.(Her Majesty’s Stationary Office, London). Retrieved 17 November 2006 from:
http://www.hm-treasury.gov.uk/spending_review/spend_sr04/associated_documents/spending_sr04_science.cfm

International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*. Retrieved 17 November 2006 from:
http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

Library of Congress 2002, *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*. (Library of Congress, Washington D.C.). Retrieved 17 November 2006 from:
http://www.digitalpreservation.gov/about/ndiipp_plan.pdf

Lord, Philip et al, 2005, *Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models*. UK e-Science Technical Report UKeS-2006-02. Retrieved 17 November 2006 from:
http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

National Science Board, 2005, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, Pre-publication Draft Approved by the National Science Board May 26, 2005. Retrieved 17 November 2006 from:
http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

OECD, 2004, *Declaration on Access to Research Data from Public Funding*. (Organisation for Economic Co-operation and Development, Paris). Retrieved 17 November 2006 from:
http://www.olis.oecd.org/horizontal/oecdacts.nsf/0/6D25ED3F86CCA316C1257088005818B2?OpenDocument

Tessella 2006, *Mind the Gap: Assessing Digital Preservation Needs in the UK*. (Digital Preservation Coalition, York). Retrieved 17 November 2006 from:
http://www.dpconline.org/docs/reports/uknamindthegap.pdf

Appendix A
Membership of the Working Group

Neil Beagrie (chair)
Richard Boulderstone (British Library)
Lorraine Estelle (Joint Information Systems Committee)
Jerry Giles (British Geological Survey)
Helen Hockx-yu (Joint Information Systems Committee)
Maggie Jones (Digital Preservation Coalition)
Michael Jubb (Research Information Network)
Chris Rusbridge (Digital Curation Centre)
David Thomas (The National Archives)
Mark Thorley (Natural Environment Research Council/Research Councils UK)
Heather Weaver (Council for the Central Laboratories of the Research Councils)

Co-opted
Juan Bicarregui (chair e-infrastructure information and data creation WG)

Virtual Membership
We are grateful also to a wide panel of expert industry, government, and international reviewers of the draft report produced by the Working Group.

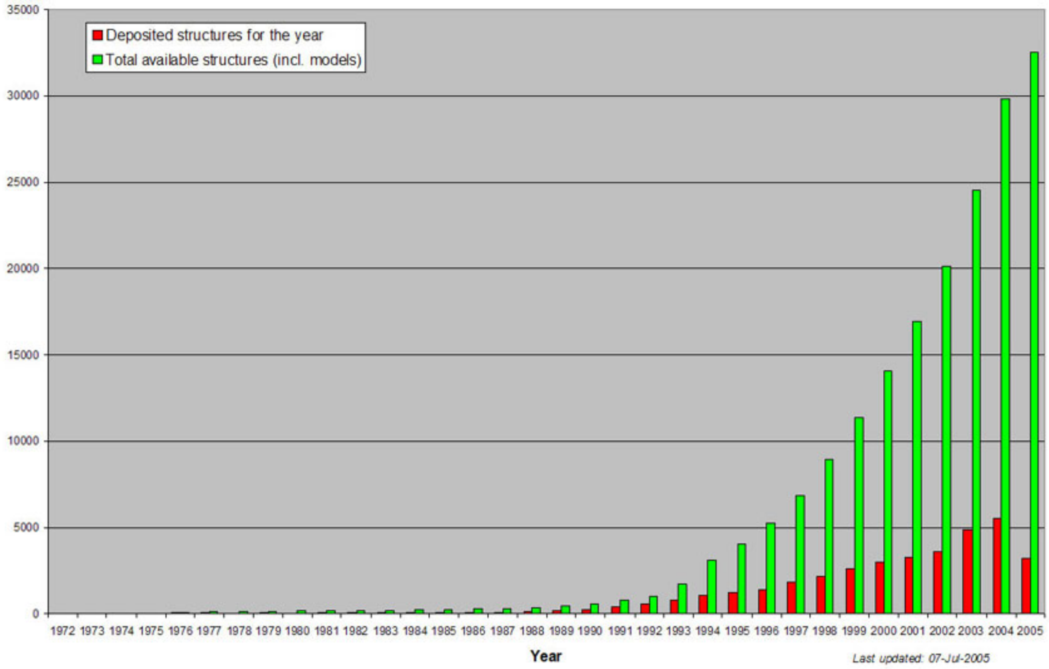
Appendix B Evidence Gathered

B1. Current Position

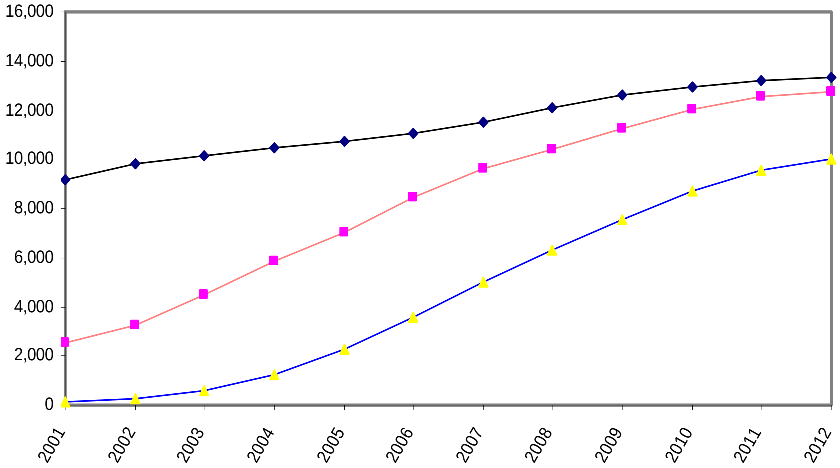
B1.1 Long-term Trends

Current concerns over developing provision reported by respondents to these surveys, is heightened by knowledge of the ongoing trends in the output of and user demand for digital research information. These trends are so dramatic that they are putting and will continue to put substantial pressure on the existing information infrastructure and working practices. Some examples for both data and publications are given below.

Examples of growth trends for research information

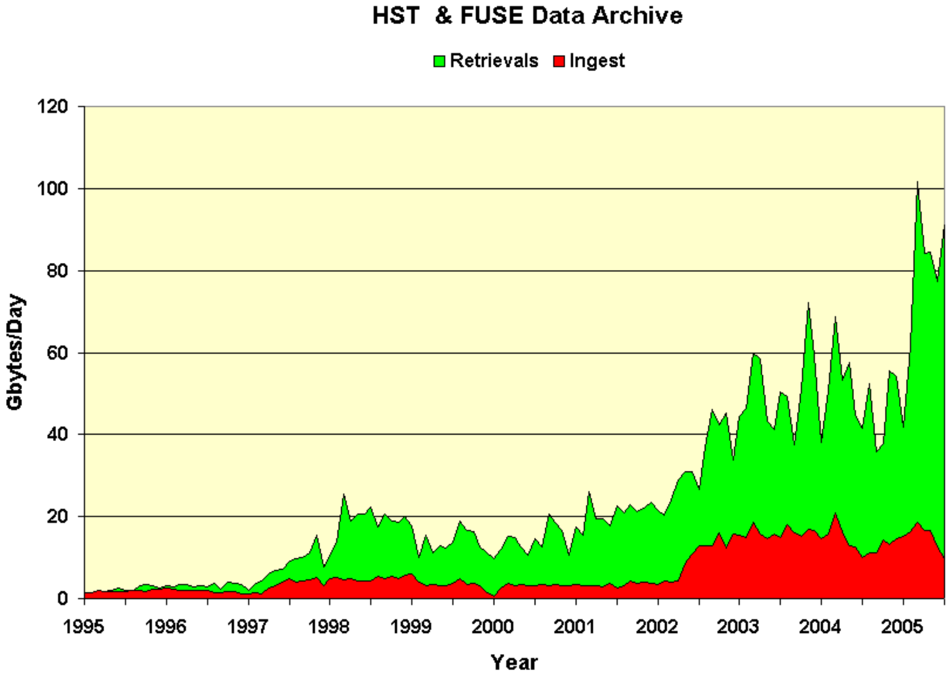


deposited structure in the Protein Data Bank 1972-2005

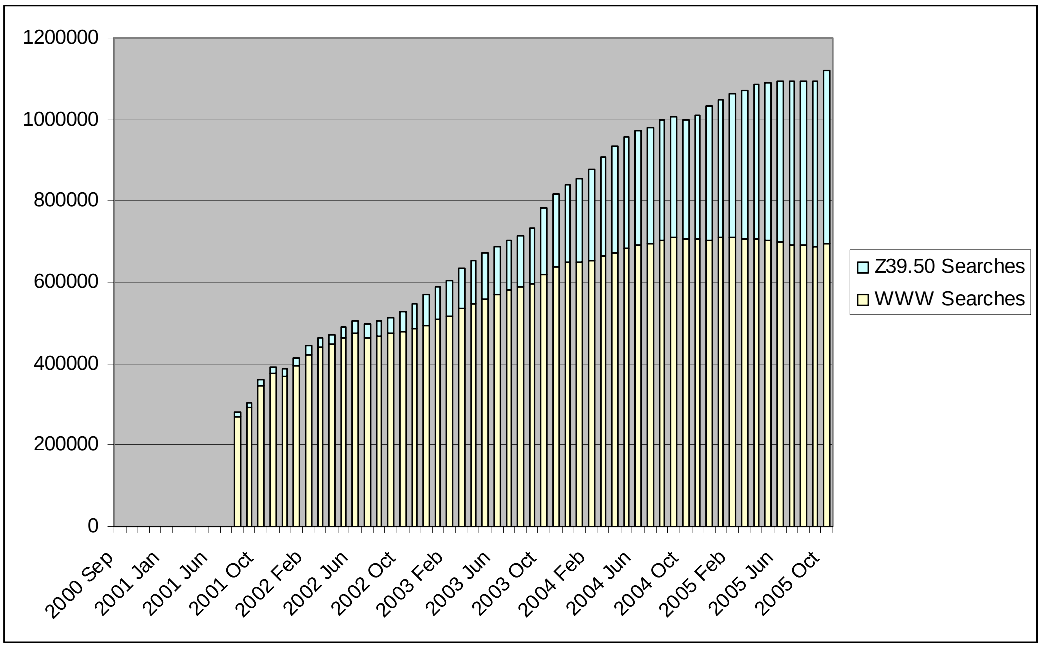


projections 2001-2012 for (from bottom to top) growth in e-only, hybrid paper + electronic, and all serial publications.

Examples of growth in user demand for digital research information



retrievals from the Hubble Space Telescope and FUSE data archive 1995-2005



use of the British Library/JISC electronic serials table of contents service (Zetoc) since launch in 2001.

B1.2 Description of the services and facilities provided

The Digital Preservation Coalition (DPC)

The Digital Preservation Coalition is a cross-sectoral membership organisation of 28 leading organisations and consortia (as of March 2006) in the UK interested in and/or involved in digital preservation. It was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base. Funding of c£150,000 per annum comes almost exclusively from membership subscriptions. It has been highly successful in advocacy and raising the public profile of digital preservation, knowledge transfer between members, identifying areas for collaborative activity, and more recently training. It also commissions surveys and reports aimed at promoting the further development of the infrastructure required to manage digital resources and to support UK research and business. The most recent example of this is the report, *Mind the Gap: assessing digital preservation needs in the UK.* This report synthesised three surveys and conducted desktop research to reveal a signficiant gap between awareness and action in the UK and identified eighteen needs aimed at improving the status quo.

Higher Education

JISC has taken a major role on behalf of the Education Funding Councils by funding a number of initiatives within its Digital Preservation programme, sponsoring a number of research reports and development projects (e-Science curation, archiving e-publications, long-term retention of and re-use of e-Learning objects and materials, requirements on preserving e-prints and the Web-archiving: a feasibility study with the Wellcome Trust). Earlier funding was provided for the CEDARS and CAMiLEON projects to investigate preservation methods.

A new development has been the establishment of the Digital Curation Centre (DCC), an initiative to provide a range of support services to the sector (and outside it) on digital curation and preservation, and to conduct research in this area. It is supplied by a consortium of the Edinburgh University, Glasgow University, CCLRC, and UKOLN. Current funding by the e-Science Core Programme and JISC takes it to 2007.

There are three university-based national data services: MIMAS at the University of Manchester holds a mixed set of data sets. EDINA is in the University of Edinburgh and also hosts a mixed set of resources, including information for the social sciences and agriculture. Also in the University sector is the National Data Repository at the University of London Computer Centre (ULCC), hosting *inter alia* the National Digital Archives for Datasets (NDAD) for The National Archives.

Many universities have set up, or are considering setting up, institutional repositories (for data and publications). Southampton and Birmingham Universities and Birkbeck College all run e-prints repositories, which allow self-archiving of research papers using the eprints software developed at the University of Southampton, which is used worldwide. Cambridge University Library and the University of Glasgow are developing institutional repositories for data based on the DSpace system from MIT, to hold a broader range of data rather than just publications.

There are many specialist repositories serving specific communities. We provide here just a small sample, taken from the field of bioinformatics:

Manchester University runs PRINTS, a detailed annotation resource for families of proteins and at the same time a diagnostic tool for newly determined genome sequences. The

European Bioinformatics Institute (EBI) at Hinxton near Cambridge is a leading centre for genomic and proteomic information, such as ArrayExpress (see also Data Collection), a repository for micro array data, the curated UniProt protein database, and the Ensembl resource, which provides access to some 20 genome sequences including man. These resources are massively used: the EBI recorded two million hits on a single day in April this year, for example. Also at Hinxton is the Sanger Centre, funded by the Wellcome Trust, providing curated genetic data and tools to explore them such as the GeneDB, which allows integrated searching of the gene databases held by the Pathogen Sequencing Unit. The Nottingham Arabidopsis Stock Centre (NASC) provides a resource of plant genomics data to a world-wide user base.

There is lack of repository provision for small research bodies, local bodies, professional groups and the general public to lodge data resources. These people are unlikely to be able to afford their own repositories or find other suitable hosts. Such repositories will require scoping and defining the nature of services to be provided, but it is a vital piece of infrastructure that will not be funded or viable without central resource. Such provision will help capture those small information nuggets that may get lost, and can play a strong role in educating people generally about e-research and could stimulate greater interest in science in the public at large.

An example of work which can be undertaken by professional groups is The Geological Society of London, the geologists' professional body and a Learned Society celebrates its 200th year next year. It is developing the "Global Connection Centre for the Earth Sciences". This will digitise the entire catalogue of the library (300,000 books, 800 current journals, and worldwide map coverage) and develop full access to the Centre's resources via the Web. This is largely being paid for by gifts from the oil industry.

Industry

It is important to note that currently the terms digital preservation and curation would not be widely understood or used across industry sectors although some of the underlying issues would be familiar and solutions to them of common benefit to industry, science and the public sector.

For example retention of electronic records for many decades is a growing issue for many industry sectors such as pharmaceuticals, petrochemical and nuclear industries, environmental science and engineering, oil and gas companies, healthcare, financials, the legal profession, and aerospace for compliance with UK and international legislation. Compliance and records retention brings in the long-term issues associated with digital preservation such as obsolescence, authenticity, and access over long time periods.

However compliance is largely seen an externally imposed cost by management and in the long-term may be less significant as a driver in this area for industry. Business is as influenced by the growing volumes of digital information, the requirement for greater inter-operability between different data sources and data types, and the need to add value to and exploit corporate information. Issues such as reducing the cost and complexity of maintaining and exploiting information through the provision of more open file formats and data exchange formats, information lifecycle management policies, and more effective mass storage are also required in similar ways as for digital preservation and curation. There are therefore shared benefits and interests. Emerging examples of this are current moves towards defining open specification standards for file exchange formats by Microsoft or the OpenDocument format (now an ISO standard), standard archival file formats by the digital photographic industry, or archival PDF formats by Adobe.

The PLANETS digital preservation research project consortium has estimated that the value of digital documents produced in the EU which are in danger of digital obsolescence – if no action were taken to preserve them – is in excess of €3 billion per year. This is based on conservative estimates for the value of documents and the decay of this value over time. The consortium also estimated that if the average lifetime of file formats can be extended by just two years, the value of documents endangered by digital obsolescence might be reduced by as much as €1 billion euros every year.

Substantial work has been done on information standards in engineering – notably on the STEP (ISO 10303) standards – and this sector is increasingly looking at medium to long-term issues for interoperability and system independence. For example there is work currently underway in the aerospace industries to produce an international standard for Long Term Archiving and Retrieval (LOTAR) of digital technical product documentation such as 3D, CAD and PDM data (AECMA 2006) utilising the ISO standard reference model for Open Archival Information Systems – OAIS (ISO 14721).

In the aerospace sector it can be argued that the major emerging digital preservation challenge is one of model retention. Product behaviour is increasingly being simulated before the physical product is made, and the design models are increasingly used as a direct input to the manufacturing machines and robots. Any particular presentation displays only part of the model appropriate to the task being performed by a user. Further, model information is interchanged both along the process chain and up and down the process hierarchy. Consequently, the proof that a part is safe to use relies increasingly simulations driven directly from design models.

Most industry sectors have reached the tipping point where electronic data is the primary source. In engineering there is a second tipping point, where the model is primary, rather than the presentation of the model. This point is already in view for the physical design of products, and as tools grow in sophistication, this will affect most areas of engineering. When this tipping point is reached, any return to non-electronic documentation will no longer be possible. It is one of the aims of the LOTAR project to ensure that engineering companies can retain digital design models over the long-term, and hence can safely cross that second tipping point.

There are two issues that arise from this new engineering environment dependent on digital modelling. Firstly, there is a practical need to retain information on the facilities and equipment used. For example, at the lowest level of risk, loss of the building CAD model will mean that the cost of projects like recabling will be unpredictable. Approval for facilities involving chemical, biological or nuclear hazards will be contingent on showing that they can safely be decommissioned, and consequently that any electronic models needed for decommissioning will be available in the long term. Secondly, for complex experiments, retaining the model of the experimental system may be as important as the results. A further consideration is that the same digital models needed for long-term retention can be used for intra- and inter-enterprise integration, and this is currently the driver for technologies and engineering standards such as STEP.

Other major industries such as publishing, media, and science based commercial research also have a vested interest in seeing the long-term preservation and curation of digital information addressed by others even if they do not undertake this activity directly themselves. For example the DTI commissioned report 'Publishing in the Knowledge Economy' undertaken in partnership with the publishing associations argued that “in order for the UK to protect access to important research material and to ensure that small and not-for-profit publishers are not unfairly disadvantaged, the archiving of digital research should be organised at a national level by Government.” (DTI 2003).

There are some outsourced/aggregated providers of industry information. Common Data Access for North Sea oil exploration data is one such, working in collaboration with NERC’s British Geological Survey to hold the National Hydrocarbons Data Archive as well as other data. Commercial providers of electronic content as well as traditional formats such as newspapers have digital repositories of their materials for internal and business to business use. Media companies (broadcasters, film, advertisers) also maintain such repositories. There are a few private digital data archives in the pharmaceuticals industry (Pfizer in the UK, GlaxoSmithKline), and media sectors. Many commercial companies have their own research libraries (e.g. the pharmaceuticals companies).

Research Councils:

AHRC	Joint JISC & AHRC supported Arts and Humanities Data Service. This service is divided into five subject-specialist units located at academic centres of excellence, with a central coordinating office at Kings College London. The AHDS provides curatorial services and extensive support for users. It is planning to centralise its storage arrangements, leaving the five units with preservation and subject-specific support functions. AHRC Joint ICT policy with AHDS covering archiving. For grants awarded where a significant product is the creation of an electronic resource, data and documentation must be offered to AHDS within 3 months of the end of the award.
BBSRC	BBSRC has undertaken a consultation on a draft data sharing policy and a formal policy is planned to be published for implementation in summer 2006. Researcher data sharing obligations are under consideration as part of policy development. BBSRC will aim to encourage and facilitate data sharing and is proposing to provide: (a) Funds to support community resources and facilitate development of data sharing approaches in specific communities and to support development of standards and software tools which enable data sharing. (b) Information and guidance to applicants including information about existing standards, guidelines, databases and resources that may be relevant. (c) Support for relevant training activities.
EPSRC	Data managed in a durable form under control of the institution of origin. EPSRC does not overly intervene in the research dissemination process and has no formal policy in this area. It encourages investigators to manage primary data as the basis for publications securely and for an appropriate time in a

	durable form under the control of the institution of their origin. A consultation with the EPS community on the research output dissemination process and common practices for data sharing is currently underway.
ESRC	Joint JISC & ESRC supported UK Date Archive, including the Economic and Social Data Service. This provides preservation services and conducts research into data management and preservation. Formal ESRC data policy. Current version dates from 2000, but to be updated soon. Applicants must carry out a data review to ensure funds are not requested for data that are already available. Data must be offered to the Data Archive within 3 months of end of award.
MRC	Encourages curation and long-term management. Onus is on the PI and their institution. MRC have recently implemented a Data Sharing and Preservation Policy that will apply to funding proposals awarded from January 2006. From this date applicants must produce a plan for data sharing and preservation.
NERC	NERC supports seven discipline based, long-term Designated Data Centres that are organised in a similar fashion to the AHDS. These data centres fulfil the role of national stewards of environmental data. The NERC Data Grid is being developed to provide a seamless access gateway into these resources. In addition, NERC supports data management activities in support of specific research programmes, including the NERC Environmental Bioinformatics Centre. NERC has a well-developed Data Policy and the current version dating from 2002 is currently being updated. NERC funded programmes must have data management plans. All data collected under NERC funds must be offered to a NERC data centre however PIs allowed reasonable time for first use.
PPARC	Supports long-term curation of selected data sets. Curation and sharing driven by the science and recognised that it is not appropriate to maintain all data. Subject specific data centres are supported as projects. The general principles of the PPARC data policy are that data generated through PPARC's research programmes, should, where possible and economic, be made available to all (after a proprietary period, if applicable). New projects are required to formalise data ownership and agree distribution mechanisms before projects are funded. Projects are required to address long-term data curation at the end of a mission or experiment. Projects are given the flexibility to adapt their data curation plans to suit the needs of science exploitation.
CCLRC	The CCLRC funds a high-capacity storage facility at RAL for archiving large data sets based on robotic tape storage. This store has a nominal capacity of 1Pbyte and has been operational since 1986 and in its present configuration since 1998. This store as yet provides limited curatorial services for specific data sets, though plans are being made to move to an increased curatorial role. CCLRC is working on several e-Science projects aimed at this.

Libraries, Museums and Archives

In contrast to the data stores summarised above these resources provide stores focused more on documentary information, sound, images and video items, as opposed to large volumes of numerical data. There are many hundreds of university and college libraries, national, regional and local libraries, archives and museums. We summarise here only the largest UK institutions (The British Library and the National Archives). Other major providers are listed in section B1.3.

The British Library is an internationally renowned resource, with a vast collection of some 150 million items in many formats, both digital and physical (including 3.5 million sound recordings). Its catalogues are online and resolve some 150 million searches per annum. It is a legal deposit library, the other five being the National Library of Scotland, Bodleian Library, Cambridge University Library, Trinity College Dublin, and the National Library of Wales. The British Library is making a major investment in long-term digital preservation through its DOM programme. The vision for this programme is to “enable the UK to preserve and use its digital output for ever”. The DOM programme is creating a secure archive primarily for text-based material but is also exploring the storage of audio, maps, and web content as part of its legal deposit remit. A digital preservation team is being setup to ensure the long-term issues of technological obsolescence are effectively managed.

The National Archives of England, Wales and the United Kingdom (TNA) has one of the largest archival collections in the world, spanning 1000 years of British history, from the Domesday Book of 1086 to the latest born-digital government records. TNA has a dedicated Digital Preservation department, and has been operating a digital archive for born-digital public records since 2003. Through its Seamless Flow programme, TNA is now developing end-to-end procedures and automated systems for managing its electronic collections, including appraisal and selection, transfer to TNA, long-term preservation, and delivery to users. The PRONOM technical registry has been developed by TNA as a service to support preservation planning. Its database contains detailed information on over 550 current and obsolete file formats, together with the software required to access them. Wherever possible, information for this database is provided by software vendors. The PRONOM Unique Identifier (PUID) scheme provides persistent unique identifiers for file formats recorded in the registry, and has been adopted as the preferred encoding scheme for file formats within the e-Government Metadata Standard. TNA has also recently released DROID, a Java tool for automatically identifying file formats, using signature information stored in PRONOM. Electronic Records Online, the pilot version of TNA’s web-based delivery system, was released in 2005. The National Archives of Scotland and Public Record Office for Northern Ireland are both also developing systems for preserving electronic records.

Both the British Library and The National Archives are partners in the UK Web Archiving Consortium, a collaboration between six leading UK institutions, led by the BL, to develop a test-bed for selective archiving of UK websites.

Unaffiliated research and private individuals

Private individuals are producing an increasingly large quantity of personal, digital information, some of which resides on their own computing devices and some externally. Email, financial records, documents, audio, images and video are key examples. These “personal digital collections” are of significant value to the individuals concerned and in the case of a smaller number of notable examples will form the basis for more widely relevant cultural and scientific memory which will become available to researchers through libraries, archives and other repositories. While facilities and services for personal archiving are beginning to appear, they remain limited to providing backup and file store functionality with no genuine long-term preservation capability. R&D in digital preservation and curation may provide a framework within which private enterprise will have the opportunity to capitalise on these activities and emerging needs for mass market long-term storage and access for the digital memories of individuals.

B1.3 Identification of the major providers

Category	Major components
K. Preservation	
	JISC Preservation development programme
	Digital Curation Centre
	Digital Preservation Coalition
	National Archives, including the PRONOM system
	British Library, including the Digital Object Management programme
	BBC

Category	Major components
A. Information Curators	
Current:	
Libraries & archives	British Library
	National Libraries of Scotland and Wales
	Other libraries of deposit (CUL, OUL, TCD,)
	The National Archives
	National Archives of Scotland; Northern Ireland (PRONI)
	National Film & Television Archive
	Specialist libraries and archives
	Local archives and record offices
	Libraries in commercial organisations
	Libraries in national laboratories, learned societies
	Research libraries in HEIs (institutional, faculty, departmental)

Category	Major components
A. Information Curators	
Data stores	Atlas Data store AHDS, five subject repositories NERC, eight thematic data centres UKDA ULCC EDINA MIMAS
Digital repositories	A growing number of institutional and self-archiving repositories using software such as LOCKSS, Fedora and DSpace JSTOR (Journal archive, UK mirror site at MIMAS)
Publishers	Commercial, learned societies, university presses, – many and various HMSO Abstracting and Indexing Services (A&I) - various
Government agencies	UK Government Departments of State and Executive Agencies UK Local Authorities Meteorological Office European Centre for Medium-range Weather Forecasting Ordnance Survey British Geological Survey UK Hydrographic Office

B1.4 Estimate of current expenditure

The research councils do not identify expenditure on digital preservation and curation as a separate budget heading except where they are funding data centres. It would be valuable to identify future metrics which could track and report other relevant expenditure.

Currently the JISC and AHRC spend c£.11m per annum on the Arts and Humanities Data Service, ESRC and JISC c£1.8m per annum on the Economic and Social Data Service, and NERC c£5.5m per annum on the 7 NERC data centres.

The UK government science research budget in 2005-6 is allocated currently as follows: Arts and Humanities £68m; Biotechnology and Biological Sciences £321m; Central Laboratories of the Research Councils £301m; Engineering and Physical Sciences £575m; Economic and Social Sciences £126m; Medicine £503m; Natural Environment £370m; Particle Physics and Astronomy £342m; other UK Gov. depts £701m. Funding for preservation and curation across the research councils varies according to needs and scale of the disciplines involved. However from the above it can be seen that where data centres and services exist they represent approx 1.4-1.5% of total research expenditure (excluding indirect overheads and full economic costs).

The JISC contributes c.£780,000 and the EPSRC (via the e-science core programme) c £500,000 per annum (c£1.28m pa in total) to the Digital Curation Centre. JISC core programme funding for digital preservation in institutions is currently £710,000 for the academic year 05/06.

In the current financial year the British Library is spending c £1.25m per annum on digital preservation. This figure covers the activities of the digital preservation team, web-archiving, and development of the national digital library storage system (DOM).

The National Archives currently spends c. £2 million per annum on digital preservation-related activities. This figure covers the activities of the Digital Preservation Department, the NDAD and web archiving programmes, and the Seamless Flow programme, which includes further development of the Digital Archive and PRONOM.

Levels of investment by Industry are largely commercial in confidence and cannot be quantified by the working group.

B1.5 Summary of plans for future provision by the key providers.

The National Archives plans to spend c. £8 million on digital preservation-related activities over the next 3 years. Over the next two years the British Library plans to spend c. £4.75 million on digital preservation-related activities. In addition, the BL and TNA will be participating in the EU-funded PLANETS project, which is expected to receive total funding of £6 million over the next four years.

JISC has £14million Comprehensive Spending Review money allocated for supporting capital development of digital repositories, digital preservation and shared services in universities over the next 3 years. It is difficult to say how much of this is for preservation as it is an integrated part of the work. JISC core programme funding for digital preservation in institutions currently £710,000 for the academic year 05/06 will reduce to c£500,000 per annum in future years.

Funding over next 2 years for research data centres and services is expected to show small increases in real terms although some will also benefit from the introduction of full economic costs in their support funding.

The JISC funding profile for the Digital Curation Centre over the next 2 years is:

AY05/06: £779,740
AY06/07: £516,242

Six biomedical research funding bodies and charities, headed by the Wellcome Trust, have formed an alliance to create the UK's own PubMed Central, a British version of the US repository of openly available peer-reviewed scientific research. Management for this is currently out to tender.

B1.6 International comparisons

This section is not intended to provide comprehensive international coverage. Countries have been chosen to provide an appropriate international selection of leading research and emerging research countries against which the UK position and proposals can be benchmarked.

<u>North America</u>

USA

The National Science Foundation's (NSF) Cyberinfrastructure Council has initiated a comprehensive strategic planning process to guide the agency's investments in cyberinfrastructure: the IT-based infrastructure increasingly essential to progress in science and engineering. The agency's plans are being developed in a document entitled "NSF's Cyberinfrastructure Vision for 21st Century Discovery". Its draft Strategic Plan (2006-2010) for Data, Data Analysis and Visualisation envisages that: “In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form, and to transform these data into information and knowledge aided by sophisticated data mining, integration, analysis and visualization tools.” It advocates creating a national digital data framework consisting of a range of data collections and managed organizations networked together. The National Science Board (NSB) has published a report on Long-Lived Digital Data Collections. It recognizes the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, the ever increasing NSF investment in creating and maintaining the collections, and the rapid multiplication of collections with a potential for decades of curation.

The Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure published in 2005 advocated additional spending of $185 million (£106 m) on data centres and $30 million (£17 m) on digital libraries for science and engineering as part of Cyberinfrastructure development. In June 2005 NSF created the Office of Cyberinfrastructure (OCI), which has a Fiscal Year 2006 budget of $127 million (£73 m). The President's budget request for OCI in FY2007 calls for $182.42 million (£104 m) --an increase of $55.3 million (£31 m), or 43.5%.

In December 2000, the US Congress directed the Library of Congress to develop and execute the National Digital Information Infrastructure and Preservation Program (NDIIPP) and provided up to $100 million (£57 m) for this purpose. The primary goals of NDIIPP are to develop a national digital collection and preservation strategy; establish a network of partners committed to digital preservation; identify and preserve digital content that is significant and at risk; and support improved tools, models, and methods for digital preservation. The Library and the National Science Foundation (NSF) have partnered to establish a Digital Archiving and Long-Term Preservation (DIGARCH) research program as part of NDIIPP, and recently divided $3 million (£1.7 m) among awarded 10 projects to undertake pioneering investigations.

In 2005 The US National Archives and Records Administration (NARA) announced it was awarding Lockheed Martin a $308 million (£176 m) contract to build a permanent archives system to preserve and manage electronic records created by the federal government. While the full system is scheduled to be completed by 2011, a functional subset of the system will be operating by 2007.

Some interesting pointers to future priorities in US government R&D expenditure are contained in the The Supplement to the US President's Fiscal Year (FY) 2007 Budget. The Supplement provides a technical summary of the budget request for the Networking and Information Technology Research and Development (NITRD) Program. The NITRD Program, now in its 15th year, represents the coordinated efforts of the many US Federal agencies that support R&D in networking and information technology. Selected extracts below:

Highlights of the President's 2007 Request
Strategic Priorities Underlying This Request
– Long-term preservation: Maintenance of and access to long-lived science and engineering data collections and Federal records

Additional 2006 and 2007 Activities by Agency
NIH: Curation and analysis of massive biomedical and clinical research data collections; tools to manage and use new databases; tools for building, integrating ontologies; software tools for visualizing complex datasets; curation tools; build nationwide support for standard vocabularies; information integration.
NASA: Continue efforts on agency-wide data exploration architecture with centralized data repository; mobile autonomous robots and intelligent systems; speech-based human-computer interaction; wind down space exploration systems projects, including team-centered virtual adaptive automation, automated design of spacecraft systems, some robotics applications, and decision support system for health management.
EPA: Tools and approaches exploring potential linkages between air quality and human health; integration of search and retrieval techniques across environmental and health libraries; evaluation and investigation of the distribution, integration, management, and archiving of models and datasets.
NARA: Advance decision support technologies contributing to high-confidence processing of large collections(e.g., collections of Presidential records).

Canada
In June 2005, The National Research Council Canada (NRC) released the Final Report of the National Consultation on Access to Scientific Research Data (NCASRD). The Report includes many recommendations including the creation of a task force, to prepare a full national implementation strategy, and the instigation of a pilot project to show the value and impact of access to research data. The Report also proposes the establishment of a dedicated national infrastructure to assume overall leadership in the development and execution of a strategic plan which includes action to stop degradation and loss of the country's research heritage. Planning continues to establish the proposed task force. Currently only the Social Science and Humanities Research Council has a policy covering the archiving of research data.

Library and Archives of Canada (LAC). LAC is one of the few national cultural heritage institutions in the world that covers and combines digital preservation activities in the library field and the archives field in one organisation. In 2002, in partnership with the Social Sciences and Humanities Research Council, the former National Archives undertook a consultation and investigation with respect to the management, preservation of, and access to, social science and humanities research data in Canada. This National Research Data Archive Consultation was conducted by a Working Group of experts in the fields of social science and humanities research and data archiving. Their report recommended the creation of a National Research Data Archive Network and outlined three options for its administration. In the 2005-2008 period, LAC will reallocate resources to support the development of a Canadian Digital Information Strategy with other Canadian partners. It is anticipated that under this strategy, an administrative approach to the creation of a distributed network of trusted Canadian data repositories will be put forward for consideration and possible future implementation.

Asia and Australasia

Australia

The Australian government has a national strategy *Backing Australia's Ability – Building our Future through Science and Innovation*. This has committed Au$8.3 billion (£3.55 billion) funding to develop the country's science and innovation base over a 10 year period 2001 to 2010. One project funded by *Backing Australia's Ability* is the Australian Partnership for Sustainable Repositories (APSR) project. APSR is a digital preservation project with the purpose of developing demonstrator repositories and supporting the continuity and sustainability of digital collections. This also includes an investigation of the ramifications of accessing and managing research data produces for and generated by Australia's grid infrastructure and an exploration of Dspace. Partners in APSR are: the Australian National University (project leader), Australian Partnership for Advanced Computing (APAC), Universities of Sydney and Queensland, and the National Library of Australia.

The National Library of Australia (NLA) has operated a Digital Services Project since 1998 and has developed an international profile in digital preservation. This project forms NLA's key infrastructure strategy to support digital preservation activities. The aim of the project is to provide a technical infrastructure for the long-term management of digital material (both born-digital and digitised, both offline and online) in order to provide long-term preservation and permanent access. The project encompasses a wide set of IT development and procurement activities to support the overall framework and systems architecture for NLA's digital repository. Digital preservation activities (development and maintenance of the digital repository and research) are mainly funded from NLA's operational budget. The Digital Services Project is supplemented by a dedicated digital preservation program which aims to develop policy and set directions for its preservation aspects.

China

The National Digital Library of China Program is supported by the Ministry of Culture and the National Library of China (NLC) to develop digital information resources from the rich collections of NLC. One of the projects in the program is the Web Archiving project (http://webarchive.nlc.gov.cn/), focused on preserving Chinese Internet resources. It is based on China Info Mall project and cooperated with Peking University to produce the Web InfoMall system (http://www.infomall.cn). It was launched in 2001, and currently holds about 1 billion pages (15 terabyte).

The National Science and Technology Library (NSTL, http://www.nstl.gov.cn/) is the core component of Chinese Scientific Information Platform supported by the Ministry of Science and Technology. It is a virtual institution consisting of the Library of the Chinese Academy of Sciences, the National Engineering and Technology Library, the Library of Chinese Academy of Agricultural Sciences and the Library of the Chinese Academy of Medical Sciences. In addition to its role in collecting scientific information and providing services to the S & T community nationwide, NSTL started feasibility studies of a collaborative network of national archives to hold perpetual copies of foreign digital publications (especially full text journals) and also considered in its strategic planning establishing a distributed network of trusted repositories for Chinese scientific research outputs.

Japan

The Japanese Science and Technology Agency (JST) started the Digital Archive Project in 2005 to create historical back-files of academic journals published in Japan. This project aims at two goals: (1) preservation of important academic intellectual heritages of Japan, and (2) further promotion on world-wide dissemination of Japanese research results. The total budget for FY2005 is approximately 650 million Japanese Yen (£3.2 m).

The National Diet Library of Japan is working on the Digital Library Project. The two main objectives of NDL digital library services for 2004-2009 are: Building a Digital Repository; this includes Web Archiving; Digital Deposit (E- Journals); Digitisation of books, etc; and the Digital Archive Portal: developing a Portal site of digital archives throughout Japan. NDL uses the term 'Digital Archives' to refer to the digital repository system for long-term preservation. The digital preservation activities are funded from NDL's daily operational budget. The total budget for 2005 is, approximately 900 million Japanese Yen (£4.4 m).

The National Diet Library of Japan has distributed a questionnaire to companies, ministries, government offices and so forth to investigate current practice with respect to the preservation of digital information in Japan. According to the survey results, almost 60 % of the respondents preserve digital information for five years or more. They recognise the importance of saving digital information, but the obsolescence of software, operating systems and hardware is challenging.

European Union

This section is selective – given our limited remit and timescale only the European Commission and one major national player in science (Germany) have been included.

European Commission

The Commission is currently seeking views on its consultation paper for i2010:digital libraries which highlights the importance of further investment in digitisation, digital preservation, and digital access. A separate consultation on i2010 and scientific information will be undertaken later this year. EU investment in digital preservation research is increasing significantly with some 18 million euros (c.£12m) allocated to this under call five for Information Society programmes in Framework 6. It is hoped a higher level of investment will continue under Framework 7.

Germany

Due to the federal structure of Germany a wide range of institutions and projects are working on digital preservation in the different federal states and at national level. National projects include:

kopal (Co-operative Development of a Long-Term Digital Information Archive) is a cooperative development of a shared central preservation system for digital publications. It is considered to be both a system (project result) and a project. Project partners are Die Deutsche Bibliothek, the State and University Library of Göttingen, IBM Germany GmbH and GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen). The project is being funded by the Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung, BMBF), for three years. In addition the four partners support the interface to the digital repository from the daily operational budgets of their own institutions. Procedures for packaging and preparing digital objects for ingest are implemented within the local digital libraries of the kopal partners. The kopal Project will be completed in 2007.

nestor: (Network of Expertise in Long-Term Storage of Digital Resources). nestor aims to develop a network for information and communication for current and future long-term preservation activities in Germany, to establish a cross-sectoral community to promote and support long-term preservation activities and to raise awareness in society, to trigger synergies between ongoing activities in Germany and to cooperate with international partners and projects, and to develop strategies for the coordination of long-term preservation activities in Germany. The project is also funded by BMBF and will end in 2006 with a proposal for a follow-up project to establish a long-term organisational model to continue the service as a network of excellence, along the lines of the DPC in the UK. nestor's partners are libraries (Die Deutsche Bibliothek, Bavarian State Library, Lower Saxony State and University Library), a media centre (Computer and Media Service of Humboldt University, Berlin), archives (Bavarian State Archive - Head Office, German Federal Archives) and a museum representative (Institute for Museum Studies, Berlin). The nestor Advisory Board consists of publishers, representatives of science & technology, museums, archives, libraries and universities and as well as members of culture & politics and research institutions/computing centres.

The German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) is also very active in funding for digital preservation projects.

Appendix C Research Agenda

C1. Ideal situation in 10 Years

Technology

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.

Where we would like to be	How to get there
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.	Further work with commercial systems providers and key service providers and user groups. Develop and standardise interfaces to allow "pluggable" storage hardware systems. Standardise archive storage API i.e. standardised storage virtualisation. Certification processes for storage systems.
"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.	Develop increasingly powerful virtualisation tools and techniques, with a particular emphasis on knowledge technologies. Develop protocols and information management exchange mechanisms, including synchronisation techniques for indices etc., to support federations. 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. Standardised APIs for applications and data integration techniques. 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. 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. 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.

Organisational, social, and economic

Over the 10 year time horizon one would need to see:

A more widespread appreciation of digital preservation. This 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.

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.

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.

Metadata standards in place at all levels, more widely deployed and implemented, and much more scaleable 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.

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.

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.
- 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.
- 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.
- 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.
- 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.
- 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
- 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.