Invest to Save

Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation

(2003)

Margaret Hedstrom, University of Michigan
Seamus Ross, HATII, University of Glasgow

Kevin Ashley, University of London Computing Centre
Birte Christensen-Dalsgaard, Statsbiblioteket (Denmark)
Wendy Duff, University of Toronto
Henry Gladney, HMG Consulting
Claude Huc, French Space Agency
Anne R. Kenney, Cornell University
Reagan Moore, San Diego Supercomputer Center
Erich Neuhold, Fraunhofer (Darmstadt)

Prepared for:
National Science Foundation's (NSF) Digital Library Initiative
&
The European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS)

Invest to Save
Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation[1]

Executive Summary

The need for digital preservation touches all our lives, whether we work in commercial or public sector institutions, engage in e-commerce, participate in e-government, or use a digital camera. In all these instances we use, trust and create e-content, and expect that this content will remain accessible to allow us to validate claims, trace what we have done, or pass a record to future generations. Unfortunately, with the current state of knowledge about digital archiving and long-term preservation there is a great risk that valuable digital content will not survive for the 'long term', even if 'long term' is defined as:

> *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. [4]*

Many organisations, businesses, government agencies, and even private citizens will need digital archiving mechanisms in order to retain access to their own records in the face of constantly changing information and communication technologies (ICTs). Cultural institutions need more reliable and more affordable digital archiving methods, systems and technologies if they are to extend their missions of preserving society's cultural heritage and intellectual capital into the digital age.

Concerns over digital archiving challenges span many types of content and a wide variety of institutional, legal, social and technological contexts:

- E-government: In Europe and North America, initiatives have been launched which aim to use ICTs to improve the way governments communicate with and serve the needs of citizens.

[1] National Science Foundation (NSF) Award #0207482 and the European Commission under IST-1999-12262 (DELOS). The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation or the European Commission.

E-government initiatives exploit ICTs with a number of goals in mind, including dissemination of information to the public, providing enhanced government services online, reducing paperwork, improving inter-governmental communications, and engaging citizens directly in democratic processes. Many of these transactions produce digital records that must be organised, managed and retained for purposes of audit and public accountability.

- E-commerce: Businesses of all types and sizes are using ICTs to cut costs and increase the size, reach and scope of markets for products and services. Typically, records of these transactions must be retained for a few years for purposes of auditing and demonstrating compliance with various laws and regulations. Some business records, however, must be kept for much longer periods of time. In the pharmaceutical industry, for example, all data associated with the development and testing of new drugs must be kept at least as long as the drug is in use.

- Education and Research: New ICTs have the potential to transform learning, teaching and research. Current technology already provides the means for thousands of online courses in basic subjects, advanced specialisations, and for continuing education. Individual instructors and institutions are investing time, money and effort to build teaching resources which, if managed adequately, could be reused by their originating institution or sold for use or repurposing by other educational providers. At the same time, researchers and scholars are collecting and creating vast quantities of research data that requires careful curation and long-term management [27]. While many of these resources are of unquestionable value for education and scholarship, preserving them for the long term remains a challenge for universities, libraries, data archives, publishers, and for the content creators themselves.

- Digital Libraries: The transition of our documentary heritage from print to digital media has led to an increasing recognition by national, academic, commercial and public libraries that to serve the needs of their current and future users they must develop digital repositories that will enable them to ingest, make accessible, conserve and preserve digital entities. Moreover, they recognise that to do this cost-effectively they must have access to tools to automate many of the processes from metadata

extraction and completeness verification to authenticity maintenance, security and migration.

- E-heritage: The cultural heritage sector is not immune from significant developments in ICTs. Libraries, archives and museums have already developed many rich online resources, often with support from government and private foundations. Early initiatives in the cultural heritage sector concentrated on converting conventional materials (manuscripts, photographs, art works and museum objects) to digital form; the heritage sector is now grappling with preserving these materials for the long term. Moreover, many new forms of cultural expression, such as art, are created in digital form and often distributed through the Web [30]. Considerable research and development is needed to improve archiving systems and technology so that cultural institutions can carry out effective and affordable stewardship responsibilities for digital collections.

- Personal Archives: More and more people communicate via e-mail, keep their accounts online, capture significant events with digital still and video cameras, and share information through their personal Web pages. Many will wish to preserve these precious materials to share them with future generations. They need digital archiving tools that are inexpensive, reliable, widely available, durable, interoperable and easy to use.

This report identifies significant research challenges in the area of digital archiving and long-term preservation. While unique problems exist in every sector from government, private companies, and schools and universities to cultural heritage institutions, our intent is to define research challenges and development opportunities that will lead to common solutions that can be shared among these sectors. What are the common requirements for these preservation capabilities that can be approached through basic research? Is there a common software infrastructure that will sustain preservation practices across cultures? Where do different types of content, different data formats, and different user communities demand unique approaches?

The digital environment is fundamentally reshaping how society is producing, disseminating, using and repurposing information and knowledge. This transformation requires effective digital archiving solutions as part of the infrastructure for a knowledge- driven economy. The attempt to replicate traditional mechanisms for appraising,

acquiring, documenting and managing information in the digital environment has not provided mechanisms that respond to the complexities and fluidity of digital entities themselves and their contexts. While acknowledging the value of many conventional archival principles, the Working Group concluded that archival processes must be redesigned and re-engineered. This change will require a paradigm shift in research if it is to provide the innovation, whether theoretical, methodological or technical, necessary to enable long-term access to digital materials.

There are two problems related to the preservation of digital entities that need generic solutions – interpretability and trustworthiness. For this to be done effectively, methods and mechanisms for automating preservation process are essential. We take as a given that technology will continue to evolve, new kinds of digital entities will be developed, user expectations both of the kinds of material with which they can work and of the ways they can work with them will rise, and the quantities and diversity of materials that will need to be preserved will keep on expanding. New research projects in digital preservation and archiving need to: (a) concentrate on a specific aspect of digital preservation; (b) focus on tangible deliverables; (c) place more emphasis on engineering and computer science that is informed by archival issues and long-term digital preservation requirements, and (d) recognise that digital entities are the raw materials for the industries of the 21st century – intellectual capital.

RESEARCH AGENDA

After examining existing and ongoing research, rapidly evolving information technologies and new application environments, the Working Group identified research possibilities in the area of digital archiving and long-term preservation. This agenda will make possible new theoretical approaches, viable tools and methodologies needed to respond to the array of challenges created by technology evolution.

We have structured the research agenda for digital archiving and preservation into three main areas: (1) Emerging Research Domains, (2) Re-engineering Preservation Processes, and (3) Preservation Systems and Technology.

In the full report we also note policy, organisational, educational and other activities that require attention and support, but which we

concluded fell outside the research and development work traditionally funded by the European Commission (EC) and the National Science Foundation (NSF). We would hope that other funding agencies at regional and national levels would support areas of research identified under this agenda. This research needs to be conducted on an international basis because the challenges are global and they do not respect national boundaries.

1. Preservation Strategies: Emerging Research Domains

1A: Repositories: There are four areas for research to support the development of repositories:

1) Elaboration of existing repository models leading to technical specifications and standards that can be used to build persistent archives. This would include development of a service layer that would allow distributed repositories to share content, tools and services. Existing models also need to be tested for scalability.
2) Software repositories. Emulation and salvage and rescue techniques rely on software that may no longer be available for purchase or licensing. The preservation community would benefit from a small number of software repositories that would collect, maintain and distribute obsolete software. Actually making the repository function effectively depends upon new research in such areas as engineering, software, systems and formalised testing.
3) Format repositories. Registries of digital formats provide keys to understanding the nature of digital objects, guide the managing of their transition from one state to another, and inform the choice of preservation method for material in specific formats [1].
4) Repositories of peripheral devices. A major obstacle to salvage and rescue, migration and emulation is the difficulty of finding peripheral devices (tape and disk drives, displays, control panels, etc). Research areas include the feasibility of engineering generic connections to enable newer hardware to communicate with legacy peripheral devices.

1B: Archival Media: To bring new classes of technology to bear on the recovery, reconstruction and interpretation of the meaning represented by bitstreams, they need to be encoded in preservation formats and on 'archival media'. Research into generating cheap, long-lasting, efficient and verifiable media for storing the bitstreams is needed.

1C: Salvage and Rescue: Preservation strategies depend upon our ability to access storage media over time. While we know that some storage media can have a shelf life of thirty years or more, the devices for reading particular classes of media tend to have much shorter life-spans, often only a couple of years. While a peripheral device repository might help here (see above), generic devices capable of reading diverse classes of media are needed to address peripheral device obsolescence.

1D: Storage abstractions: Preservation systems map between the operations that can be done on digital entity encoding formats and the operations that are supported by storage repositories. As newer classes of storage devices are developed research will be necessary to identify how their emergence will change digital entity encoding formats to take advantage of content-based addressing and parallel processing of data. Holographic storage is an example of a format that will require this kind of research.

1E: Documentation of Functionality and Behaviour: Preserving both digital entities and their underlying technologies depends upon representing their functionality and behaviour. This research should lead to the development of an extensible formal descriptive language for the performance and behaviour of preserved digital entities that would allow future users to measure how far the performance or behaviour of a digital entity deviated from its original performance.

1F: Context-aware Digital Entities: Increasingly research into agents and self-awareness among digital entities and systems has demonstrated a rich array of possibilities. We recommend digital archiving research that focuses on context sensitivity, risk awareness and proper preservation behaviour.

1G: Accelerated Ageing: Conservation of analogue items benefits from research into how materials age. There is room for new research into the area of the accelerated ageing of media, systems and software, aimed at predicting the risks to digital objects caused by software obsolescence, changes in standards or product failures in the market, rather than the ageing of physical objects.

1H: Accumulation and Preservation of Intellectual Capital: The concept of preservation is being extended to include preservation of the knowledge inherent in digital entities and the processes used to create them. For some communities, the ability to analyse the information and

knowledge content of digital entities is the most important aspect of preservation systems. This raises complex issues of the semantics necessary to represent temporal, procedural and spatial relationships and the means to relate these relationships to digital entities.

2. Re-engineering Preservation Processes

The expense of current approaches to digital preservation reflects the significant amount of human intervention they involve. Where preservation processes can be automated these costs can be reduced. New research is needed to establish mechanisms to identify processes that can be automated and to develop methods to automate them.

2A: Modelling Preservation Processes: With some exceptions, such as preservation systems needed for digital libraries, a growing body of opinion indicates that the preservation of digital entities can be enhanced if preservation functionality is built into the digital entities or the systems that manage them at the time they are created. This means improving our knowledge about what preservation functionality really is and ensuring that this functionality can be effectively communicated to system developers, modelled and implemented by them.

2B: Automation of Processes: The preservation of digital entities depends upon active curation. Human intervention at each stage of the preservation process is not economically viable. Processes that can be automated need to be identified and mechanisms for automating them developed. For example, what are the particular capabilities required to automate the processes of appraisal, accessioning, description, arrangement, preservation and access of digital entities?

2C: Detecting Trustworthiness and Information Quality: Belief in the integrity and authenticity of digital entities underpins their possible re-use and the weight that they will be given by eventual users, whether human or machine. Tools are needed to enable future users of digital entities to determine whether they have these qualities.

2D: Scalability: With a few exceptions preservation research to date has involved work with small sets of digital entities. As a result, the costs and efforts associated with larger collections have not been effectively benchmarked. At the other extreme, can we develop inexpensive preservation tools and technologies that individuals without extensive archival or IT skills can readily use?

2E: Collection Completeness and Anomaly Detection: Users need information about the completeness of a collection. Methods and tools for providing this data are currently lacking. Is it possible to detect when collections are incomplete? How can the completeness and closure of collections be validated as part of the accessioning process? Is it possible to differentiate between anomalies or artefacts and inherent knowledge within a collection that has not been expressed?

2F: Distributed and Grid Storage: Newer storage strategies offer the potential to reduce risk through the distribution of content across a network of devices. What impact does storage of this kind have on the naming, management, discovery and delivery of digital resources?

3. Preservation of Systems and Technology

3A: Formats of Digital Entities: Digital entities consist of complex objects including audio, moving image material, data held in databases, Web pages. Insufficient research has been directed at developing preservation strategies and standards for emerging digital formats, such as digital audio, digital video, models and simulations. Projects are needed that address the specific characteristics of a wide variety of formats beyond text, data and images.

3B: Managing Complex and Dynamic Digital Entities: Many digital entities are dynamic, meaning that they change as a result of adding new data or interacting with other digital entities. For example, dynamic documents are increasingly dependent upon data that might have variable instantiations and be held in databases and spreadsheets. There has been little research on the methods, tools or technologies needed to preserve these types of dynamic entities.

3C: Automated Metadata Creation: The creation of metadata consumes substantial human resources, but it is widely acknowledged to be crucial to the long-term preservation of digital entities. How can the creation and authoring of metadata be automated?

3D: Long-term Metadata Viability: To date emphasis has been on the definition of metadata elements, but there has been limited evaluation of the effectiveness or cost of metadata for managing digital entities over time. We need research that demonstrates the value of metadata for specific purposes and the minimum amount of metadata necessary. Tools are needed to track the provenance of metadata schema, for version control, and for navigation between current schema and the schema used when the digital entity was created.

3E: Multilingual Entities and Technology: Research in the area of digital preservation has barely paid lip service to the challenges poses by multilingualism. This is not just in terms of the digital entities themselves, but also in terms of the underlying metadata, applications, documentation and user interfaces.

3F: Acceptable Loss: Under many circumstances it will not be feasible either technically or financially to retain all the functionality of digital entities and their underlying technologies. A certain amount of loss of functionality, context and meaning is to be expected. We need methods to assess the impact of preservation strategies on information loss and to inform future generations about any known information loss.

3G: Repurposing: eContent industries are recognised as fundamental to the emergence of new industries and economic development in the 21st century. The process of repurposing is generally manual, poorly understood, and not always responsive to emerging markets, which often result from the unanticipated re-use of intellectual capital.

CONCLUSION

All the areas of research described here will produce results that will have a significant impact on the efficiency and effectiveness of digital preservation. The Working Group agreed that three research areas were likely to have the greatest impact:

- self-contextualising objects;
- metadata and the evolution of ontologies, and
- mechanisms for preservation of complex and dynamic objects.

If the research options outlined in Sections 1 to 3 were to be prioritised these three should be rated highest. In the context of the value of digital assets to society's memory and heritage, its intellectual capital preservation and its future economic growth, the Working Group concluded that if we invested in focused research now we would reduce the financial impact likely to be posed by our need to access digital entities in the future and would provide an environment to promote new content-driven and creative industries.

INVEST TO SAVE
Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation

Introduction

This report presents a research agenda for digital archiving and long-term preservation developed by a joint working group, supported in North America by the National Science Foundation's (NSF) Digital Library Initiative and in the European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS).[2] The Working Group (see Appendix A) identified critical problems and issues that need to be solved so that our society has the ability to preserve valuable digital information resources for future reuse. The report sets priorities for research, discusses plausible research strategies and methods, and identifies areas where collaborative research between European and North American researchers could prove productive.

The report consists of six sections: 1) a discussion of digital preservation challenges; 2) a summary of the benefits of digital preservation; 3) a set of principles and assumptions; 4) an analysis of work to date; 5) the research agenda; 6) recommendations for additional support requirements, technology transfer, and commercialisation scenarios.

1.0 Digital Preservation Challenges

Imagine future generations attempting to understand the changes to society introduced by the evolution of the Internet: the emergence of electronic trade, the creation of virtual communities, chat, new forms of digital art, global communications, the breakdown of some twentieth-century business models, and the evolution of new social spaces and behaviour. Imagine families in the future wanting to study their heritage, trace the pathway of a genetic disorder, understand why a certain decision was taken, or prove a legal right to property. All of

[2] National Science Foundation (NSF) Award #0207482 and the European Commission under IST-1999-12262 (DELOS). The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation or the European Commission.

these endeavours require continuing access to a diverse array of digital content.

Governments, businesses and organisations of all sorts in Europe and North America are moving beyond first-generation Internet-based applications and embracing concepts of e-government, e-business, e-science, telemedicine and online learning. These new applications promise fundamentally to transform commerce, education, research and interactions between governments and citizens. The next generation of applications will utilise the rapidly expanding computational and communications infrastructure to combine highly curated and well-managed digital resources with powerful tools for information processing and analysis. They hold great potential for improving access to a wide variety of cultural heritage resources.

Unfortunately, with the current state of knowledge about digital archiving and long-term preservation there is a great risk that valuable digital content will not survive for the 'long term', even if 'long term' is defined as:

> *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository*. [4]

There are severe limitations to current methods, processes and preservation strategies, systems and technologies. All current and proposed approaches are labour intensive, do not respond to business, cultural and professional obstacles to digital preservation, are heavily dependent upon continued awareness on the part of different players to knowledge about the preservation requirements of the digital entities in their care, have not demonstrated suitable levels of scalability, and have not been applied repeatedly to the same digital entities.

Digital preservation, with the *de facto* central role of technology, reflects the shifts in the broader technological realm from hardware to software, from mass production to responsive services, from simple to complex objects, from storage to functionality. If technology were the only aspect to address in developing solutions, the problems associated with long-term use would be – and would have been – more readily solvable. But digital preservation research is not solely a technical issue. It is very much one that encompasses organisational, legal, cultural, social and financial dimensions. The fundamental goals of preservation are also in question: what are we preserving? for whom? and why? Moreover, the

goals of digital preservation may rest in the eye of the beholder, and there are many stakeholders from content creators to users whose interests must be considered.

There are two problems related to the preservation of digital entities that we need to solve generically – interpretability and trustworthiness. For this to be done effectively, methods and mechanisms for automating preservation processes are essential. At the same time, digital preservation issues cross many different domains and disciplines, from large scientific data sets, to digitised cultural heritage collections, to Web archives, to personal digital photos and videos. Beyond generic solutions, research is needed to address the specific preservation requirements of different formats of digital content (data, text, images, audio, etc.) and the needs of different producer and user communities.

We take as a given that technology will continue to evolve, new kinds of digital entities will be developed, user expectations both of the kinds of material with which they can work and of the ways they can work with them, and the quantities and diversity of materials that will need to be preserved will keep on expanding. New research in digital preservation and archiving needs to:

(a) focus on delivering results in narrow domains;
(b) focus on tangible deliverables;
(c) have much more emphasis on engineering and computing science than it has in the past, and
(d) recognise that digital entities are the raw materials for the industries of the 21st century – intellectual capital.

2.0 Benefits of Long-term Digital Preservation

Preservation systems provide benefits to society if any of the following conditions exist:

- if unique information objects that are vulnerable and sensitive and therefore subject to risks can be *preserved and protected*;
- if preservation ensures long-term *accessibility* for researchers and the public;
- if preservation fosters the *accountability* of governments and organisations;
- if there is an economic or societal advantage in *re-using information*, or
- if there is a *legal requirement* to keep it.

2.1 Protection and conservation of cultural memory

Cultural heritage institutions such as libraries, museums and archives have traditionally played a central role in the preservation of information objects. They accept preservation as one of their core business activities and have already built considerable collections of digital assets by digitising physical objects (e.g., artefacts, documents, paintings), converting analogue audio-visual material, and gathering digitally born documents and art. The availability of appropriate preservation systems and technologies is crucial to ensure access to cultural heritage for future generations.

2.2 Global access to open knowledge and support for cross-disciplinary collaboration

Repositories of digital data have already had a significant impact on the development of international scientific and educational collaborations. Increasingly there is a trend towards data-driven science which depends not only on analysing data provided by new experiments, but more and more on re-using data sets collected as a result of earlier experiments or observations and reanalysing the data in new ways [27, 29]. Furthermore, combining long-term preservation systems with emerging WWW knowledge access and exploitation technologies will enable digital archives, libraries and open research centres to take on central roles in the knowledge-driven economy. Persistent archives thus become the resource nucleus for international education and research as well as for collaboration in the sciences, arts and humanities for both experts and students.

2.3 Preservation for accountability

Regulated industries, financial institutions, hospitals and clinics, and public entities are legally obliged to keep records over decades for purposes of accountability, continuity of operations and organisational memory. All of these institutions have a vital interest in affordable preservation methods and systems to protect their core institutional records. Even if such records are not permanently archived, many have to be kept long enough (often more than 50 years) for there to be concern about the impacts of changing technologies. A topic of extreme relevance is preservation of authentic records. The likely success of e-government initiatives will depend upon their ability to demonstrate to the citizen that digital records are maintained over time in systems that guarantee their authenticity and integrity.

2.4 Reduction of costs by information re-use

The main economic benefits of long-term preservation depend on the re-use of stored digital information that is impossible to reproduce or far too costly to regenerate. For example, research institutes such as space agencies, nuclear-physical or geo-physical institutions, or social empiric field research institutes preserve scientific and sociological empirical data for secondary evaluation, re-purposing as research problems change, and to enable longitudinal studies of change over time. Manufacturing, pharmaceutical and chemical industries make use of their technical and product documentation for further research and development. The multimedia industry and broadcasters reuse digital products to create new works.

2.5 Foundation of a knowledge economy

Advanced countries are shifting from an industrial economy to a knowledge economy. In a knowledge economy, repositories of digital information and the tools to mine, analyse and repurpose them represent a society's intellectual capital. Effective and affordable digital preservation strategies and systems will move archives into resources one can mine and reuse, and transform them from liabilities to assets.

2.6 Development of digital libraries

The President's Information Technology Advisory Committee (PITAC) report *Digital Libraries: Universal Access to Human Knowledge* recognised that digital libraries developments were essential if all citizens were to 'participate in and fully benefit from the Information Age' [24]. Libraries are essential custodians of our documentary heritage and will continue to play this role [15]. As the information- and knowledge-driven economy becomes more pervasive the role of digital libraries will become more central to its workings.

3.0 Principles and Assumptions

The Working Group developed a series of principles and assumptions to guide the development of the research agenda, set priorities, and identify appropriate research methods and project types. We were cognisant that there is much relevant related research under way, but little that focuses explicitly on archiving and long-term preservation.

3.1. The most distinctive characteristic of digital preservation is its long-term perspective. The focus on long-term preservation of digital objects makes digital archiving unique and sets it apart from

related research on information retrieval, digitisation, architecture, interoperability – even storage, security and authentication. Data management efforts from the digital library community have focused on organisation and discovery. Data management efforts from the Data Grid community have focused on federation of distributed storage resources. The preservation community needs to address issues related to data management during evolution of the underlying technologies, organisational infrastructure, and changing research interests and needs.

The timeline for action also creates a sense of urgency. We are accustomed to having a generation to decide whether we want to preserve given physical material. For digital objects the timeline is much shorter – preservation should be initiated at the time systems for generating digital materials are designed (see below) or at least when the material is generated, and certainly before the supporting hardware, media and software become obsolete.

3.2. Authenticity and integrity are core requirements. Special focus must be placed on ensuring trust in preserved digital objects. Not only are they easy to alter, but preservation strategies such as migration and normalisation involve transformations of the original bitstream. There are many possible approaches to ensuring authenticity (encryption, time/date stamps, reversible transformation, audit trails, controlled custody, replication, trusted repositories, etc.) None of these methods has been adequately developed, tested or evaluated for effective use across different formats, scalability, long-term viability or cost.

3.3. Scalability is essential for digital preservation. Preservation is challenged by the increasing scale and complexity of digital resources. Some of the scalability issues include the ingest rate (whether data can be accessioned into a repository as rapidly as it is being generated); the bandwidth and processing capability necessary to migrate an entire archive; and the repurposing rate (the capability to generate meaningful responses to user requests). Archival processes need to be redefined, formally modelled, and automated to achieve sufficient scale. Complexity of digital objects will have a significant impact on costs and scalablity.

3.4. Preservation is a continuous and dynamic process. In contrast to physical objects, digital objects cannot be put away and stored at constant humidity and temperature, and then accessed hundreds of

years from now. Preserving digital objects calls for constant attention – e.g. methods need to be implemented to ensure bits can be correctly interpreted – and that the interpretation mechanisms are documented, and that scalable methods for migration and documentation of the transactions are employed.

3.5. Preservation is done within the context of a lifecycle. Digital preservation is a concern that begins at or before the point of information creation. Research is needed at all stages of the lifecycle (design, creation, selection, organisation, storage and data management, discovery and access), and it should not be governed by present-day approaches to preservation of physical materials. The future might call for radically different approaches.

3.6. Digital preservation requires shared responsibilities. Until now, preservation responsibilities have rested with institutions like archives, museums and libraries where selection was governed by institutional collecting policies, government archives legislation, and legal deposit requirements closely tied to national systems. Today's digital resources span national boundaries and interests and challenge the distinctions between libraries, archives and museums. In the future political and legal questions about what to preserve will have to be linked to technical questions about how to preserve digital content. The amount and the nature of the data require new models for how to share responsibility for preserving the bits. Models have been suggested but their value for long-term preservation needs to be assessed. Key issues include costs and who will pay.

3.7. Multiple approaches are needed. Digital preservation research is multi-faceted and will require collaborative efforts that cross national borders, disciplines and research approaches. Given the range of needs and priorities, and the evolving and expanding set of target technologies, researchers will need to utilise a variety of methods, both qualitative and quantitative, in support of exploratory, conceptual, experimental and demonstrable approaches. In spite of the lingering hope for a 'silver bullet', experience has shown us that no single research project or group of projects will 'solve' the digital preservation problem. As the scope and priorities of the domain evolve, research in this area will require ongoing commitment on the part of funding bodies – a commitment that encompasses support for multiple research threads and tracks working on multiple levels, and encourages cross-fertilisation from other domains.

3.8. Digital preservation requires multi-disciplinary research teams. From the start, digital preservation has been multidisciplinary, drawing to it archival and information technology specialists; but it has become clear that broader input from computer scientists and engineers is required. Legal, cognitive, sociological, business, organisational and other expertise should also be tapped by and accessible to the domain. Therefore, teams of researchers with divergent backgrounds and research expertise (e.g. archival science, computer science, information studies, organisational behaviour and human computer interaction) are needed.

3.9. Digital preservation research does not stand in isolation from practice. Projects must build upon existing research and contribute to future practice. Considerable funding and effort should be devoted to the transfer of research results into organisations with digital preservation needs, technology transfer, and support for the commercialisation of pilot and prototype systems.

3.10. Preservation is a high-priority research area. There is an urgency to find practical solutions that can be applied now, but this demand must be balanced by the need to avoid quick fixes that defer, without resolving, the fundamental requirement to carry digital materials forward in a coherent, consistent, appropriate, authentic and affordable manner. Because technology changes at a rapid pace and often in unpredictable ways, it is difficult to forecast how long an underlying technology will last or how long a particular suite of digital materials can be maintained before intervention is necessary. Digital preservation research needs to shorten the cycle from identifying the archival implications of new technologies to harnessing that information to meet evolving requirements.

Digital preservation is a public interest and should be considered a priority research area by government funding bodies. Information is a vital national and international resource and forms a strong underpinning to democratic societies. Because so much information is currently generated in digital forms and because the means and will to preserve that information over time are lacking, research in this area is vital. It is essential that governments, acting on behalf of their citizens, sponsor such research.

National funding agencies in North America, Europe, Asia, Australia and elsewhere are supporting research that directly or indirectly

addresses the digital preservation domain. There are also recent examples of internationally funded research, including joint programmes between the US National Science Foundation (NSF), the UK Joint Information Systems Committee (JISC), the Deutsche Forschungsgemeinschaft (DFG), and the European Union (EU), the Social Science and Humanities Research Council (SSHRC) of Canada, and the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) [6, 17]. If even the current limited research investment is to be maximised then a commonly agreed research agenda to inform and co-ordinate this work is essential. This must be supported by formal commitments to provide ongoing support and standard measures for assessing impact and take-up of research outcomes.

4.0 Work to Date

The research into digital preservation completed during the past two decades has raised awareness of the problem and encouraged the development of policies and procedures to improve the longevity of digital resources ('It's about time'). Enhancements in metadata practices, increased use of standards, widespread recognition that intervention needs to come early in the life of digital entities, and the development of preservation models all contribute to an evolving set of best practice. Nevertheless, it has taken many years for the problems posed by digital preservation to be taken seriously by all concerned, and during that time funding, and hence research activity, has been sporadic, low-level and lacking in strategic co-ordination at national and transnational levels. In the meantime, digital content is being created at exponentially accelerating rates. Digital technologies continue to move into spheres of human activity in which it was not previously an issue, and content appears in forms that pose increasing preservation challenges.

Research and knowledge to date have derived from two principal sources. The first is funded research programmes, typically within an academic institution that leads to one or more publications or other tangible research outputs. In some cases, these outputs have also taken the form of freely available software, demonstrator systems, and archive and digital library services [1, 5, 21, 31]. Some work has also led to the production not just of academic papers but also works such as guidelines, training materials and handbooks intended to be used

within real working environments to address today's preservation problems at or before the time digital materials are created [2].

The second source has been practical experience gained by organisations faced with pressing digital preservation problems arising from their primary business objectives. These organisations include commercial concerns (the pharmaceutical and seismic survey industries, for example); government at supra-national, national, regional and to some extent local level; publicly funded scientific research organisations (such as high-energy physicists, weather forecasters, archaeology research bodies); national libraries and archives, research libraries; and organisations such as national broadcasters. The results obtained and knowledge gained through this practical experience have not always been disseminated as widely or in the same ways as formally funded research. Even where it has been disseminated, its general applicability is not always clear.

One consequence of this approach to research to date is that the products tend to be either high-level abstract models and guidelines or very specific solutions tied to a particular format of material or institutional context. The OAIS reference model, for example, offers a vocabulary and a high-level model for archival repository design, but it does not provide a blueprint for building persistent archives. At the other extreme, the BBC in the UK, for instance, has developed techniques for preserving digital TV. We can assume that these techniques will be relevant to broadcasting organisations and film and video archives in other countries. What is less clear is whether any of the experience gained is of use in other types of digital content.

There is a tension in digital preservation research between generalisable principles, methods and technologies that cut across formats, content areas, academic disciplines, and institutional settings, and the very specific requirements of different producers, content types and user communities. Organisations facing immediate and pragmatic concerns usually will do enough research to meet their current business needs, but typically will not be motivated to analyse their work for its wider relevance or applicability. We cannot determine the applicability of proposed digital preservation solutions beyond their original target unless this research is put into the public domain and funding is available for analysing its general utility.

Much early work on preservation concentrated on preserving the bitstream, or extending the useful life of the physical entity on which

information was recorded. As a result, there is a large body of work and an excellent pool of knowledge on the survival properties of a wide range of digital recording media. But new forms of information storage continue to appear, and subtle changes in formulation or manufacturing techniques can have marked consequences for the long-term survival of some digital media, even if they do not impact its short-term performance.

Work that focused on the preservation of bitstreams (rather than the media they were recorded on) has addressed issues such as scalability, management techniques, media migration and cost models. Although much of the initial work emerged from the industrial and scientific community, and hence focused on large-scale problems within large organisational contexts, it is of widespread if not universal applicability to other communities and problem domains.

With the preservation of bits beginning to seem a manageable problem, research began to focus on format or information preservation. Instead of dealing with digital information at its lowest level, this considered what information was encoded, what was significant in the long term, and how to represent that information in a way that ensured accessibility over long periods of time. This represented a shift in thinking from data preservation and data management to information preservation and information management.

Some work in this area has been very general, and has produced models for repositories or preservation processes, schema for preservation metadata, or general guidelines for content creation. Other work has been extremely format-specific. As a result, the once-complex problem of preserving static documents or structured numeric data is now well researched and well understood. In most areas, however, new formats are continually introduced, particularly with increasing use of digital storage for images, sound, video and executables, such as computer games and models. These formats are often proprietary and subject to forms of protection that greatly increase the preservation challenges. Little or no work has addressed this area.

The preservation of born-digital multimedia objects and documents, such as videos, computer games, simulations, interactive Web sites or 3D presentations, turns out to be especially urgent and problematic. The same is true for the large amount of originally analogue audio-visual content like audio, video and images transferred into digital form and

transformed to multimedia objects. Complex multimedia objects can not be expressed in traditional hard-copy or analogue media as a preliminary solution for preservation needs. More than text or image records, they are characterised as processable units and depend on the availability of adequate processing technology. But even if today's processing technology is saved for the future, eventual users will not be trained in how to use it and might even be unwilling to use obsolete technology. There is – at least – a double problem of preserving data and preserving technology.

Most research has focused on three main digital archiving strategies: normalisation of digital content into a few common formats; migration of data from obsolete to current computing platforms and applications; and emulation of obsolete platforms on current computing platforms [9, 10]. This research demonstrates the technical feasibility of these approaches, but little work has been done on issues of scalability, cost comparisons, or the effectiveness of the various strategies on different formats of digital content. There is ample room for research that explores entirely new approaches.

Little research has been conducted on complex entities. Almost all digital objects are now complex amalgams of a variety of information types. Consider a Web page that contains text, image, metadata and a small program, or a spreadsheet that interacts with a database to produce real-time reports and also contains embedded spoken and written annotations. The challenge here is not simply to preserve each component (speech, text, image, database, program), but to preserve the whole and the interrelationships among the parts in ways that ensure accessibility in the future when models and methods of human computer interaction will be very different from those we experience today.

5.0 The Research Agenda

We have divided the research agenda into three sections. The first section, Preservation Strategies – Emerging Research Domains, identifies new problems that result from constantly evolving technology. The second section, Re-engineering Preservation Processes, discusses ways in which archival processes might be re-engineered to dramatically reduce the costs of long-term data management and preservation. The third section, Preservation Systems and Technology, discusses tools and technologies that are needed to support these process transformations. The recommendations for research and

development in the second and third sections draw attention to methods, tools and technologies that could be developed and deployed in the near term to facilitate digital preservation. The emerging research areas are new issues that we identified as needing urgent attention.

Some of the research we identify is urgent and will produce benefits in the very short term (see Sections 2 and 3 below), but some research areas, including that described in Section 1, are still so young that, although promising, they will only bear fruit after several rounds of substantial funding.

5.1. Preservation Strategies: Emerging Research Domains

This section discusses new areas for innovative research and strategies.

5.1A Repositories: Repositories such as data archives provide fundamental resources to future researchers. In the area of repositories we identified four areas for additional research and development:

> *1. Elaboration of existing repository models.* Current models for archival repositories provide a useful starting point for building persistent archives, but many lack precise technical specifications. Further research is needed to develop technical specifications and standards that can be used to build persistent archives. This includes development of a service layer that would allow distributed repositories to share content, tools and services. Models and specifications for discovery, access and retrieval across diverse repositories and collections also need attention. Existing models also need to be tested for scalability. Some of this work is already under way and merits continuing support.
>
> *2. Software repositories.* Emulation and salvage and rescue techniques rely on software that may no longer be available for purchase or licensing. The preservation community would benefit from a small number of software repositories to collect, maintain and distribute obsolete software. A software repository may hold a characterisation of the capabilities of prior systems, which can be implemented using modern technology, or it may hold routines that can migrate obsolete encoding formats to modern encoding formats. Actually making the repository

function effectively depends upon new research in software and systems engineering and in formalised testing. The definition of automated testing sequences is needed to determine whether software continues to function and behave as it was originally designed to. There is also a need to address the architectural requirements for sharing obsolete software and intellectual property obstacles.

3. Format repositories. While software repositories would enable long-term access to digital assets, format registries provide keys to understanding the nature of digital objects, guide the managing of their transition from one state to another, and could be used to provide tools to analyse digital objects to determine their characteristics and to assist in the selection of the best preservation method.

*4. Repositories of peripheral devices***.** A major obstacle to salvage and rescue, migration and emulation is the difficulty of finding peripheral devices (tape and disk drives, displays, control panels, etc.). Research areas include the complexity of engineering generic connections to enable newer hardware to communicate with legacy peripheral devices.

5.1B: Archival Media: Media are increasingly robust. The newer generations of media are made with more stable materials (e.g. phthacyanine dyes in CD-Rs and very high coercivity in magnetic media) with viable life spans of upwards of thirty years under archival storage conditions. This suggests that in the future media will survive with their data intact, but we will lack mechanisms to access and interpret their content. The future will provide new technical opportunities for extracting these bitstreams, but to bring new classes of technology to bear on the recovery, reconstruction and interpretation of the meaning represented by bitstreams, they need to be recorded on 'archival media.'

5.1C: Salvage and Rescue Bitstream preservation has been shown to be viable (even if it could be enhanced by archival media), but little work has been done on developing techniques to enable raw data streams to be analysed and the original meaningful (e.g. logical) units they represent reconstructed. The use of methods such as crypto-analysis to facilitate the meaningful reconstruction, interpretation and presentation of bitstreams recovered in this way may have promise, but even small-scale experiments in this arena have yet to be conducted. We also

suggest that research into generic reading devices could provide a tool to ensure the material continues to be accessible.

5.1D: Storage abstractions: Preservation provides the ability to characterise digital entities independently of underlying software and hardware infrastructure. Equivalently, this implies the ability to create infrastructure-independent representations of digital entities. Sometimes there is no choice, e.g. one may or may not decide to collect a flash animation – but there is no choice of other representations. At present this means that the preservation strategy has to provide an infrastructure-independent representation mechanism for archived digital entities. For other media types, e.g. pictures, video or sound, there is a choice – but it will still be intimately related to the logical preservation strategy.

5.1E: Documentation of Functionality and Behaviour: We have generally adequate ways to represent static documents, databases and digital images. There are no formal ways to express the functionality and behaviour of digital entities. These are needed to establish benchmarks and measure consistency of performance across migrations or emulations. Approaches to functionality and behaviour abstraction and representation are also needed to enable us to reconstruct systems. What tests can be developed to verify automatically whether or not the system behaviour and functionality match that which the application had originally, or to document how the preserved entity deviates from the original?

5.1F: Context-aware Digital Entities: Self-aware digital entities have not received sufficient attention. If we are to ensure that preservation systems are automated as far as possible, self-awareness traits are essential. Digital entities that know what they are (e.g. tiff files) can observe the state of other objects (e.g. observe the decline of numbers of similar types or classes of objects), know where they are and where their metadata are, and can communicate with their originator or manager if they need to be protected, migrated or secured or can seek out appropriate services. These approaches would underpin self-archiving strategies.

5.1G: Accelerated Ageing: We have yet to develop accelerated ageing tests for predicting the longevity of digital information. We can test previous generation formats to evaluate their effective use in the present, but we cannot know if those findings will hold in the future until we get there. Research is needed to develop formal methods for

predicting the viable life of specific media, formats, software applications and standards. Such research must consider not only technical attributes, but also market penetration and producers' business strategies.

5.1H: Intellectual Capital: If archives are to become resources that anyone can mine and reuse, we need preservation strategies that move from the current state of data and information preservation to preservation of knowledge. Preserving intellectual capital requires much more robust ways to describe digital entities and relate them to their larger organisational, technical and domain context. This raises complex questions of the semantics necessary to describe temporal, procedural and spatial relationships and the means to relate those relationships to digital entities.

5.2 Re-engineering Preservation Processes

Most research on digital preservation has attempted to map traditional preservation practices for physical materials onto digital content. It is increasingly clear that digital preservation strategies require a complete re-engineering of digital preservation processes. Ideally, re-engineering will produce new processes that can be highly automated, reserve human judgment for issues where it is most needed, and produce approaches to preservation that are cost-effective and scalable.

5.2A: Modelling Preservation Processes: Preservation of digital objects can be enhanced if preservation functionality is built into systems used to create and manage them. This means improving our knowledge about what preservation functionality really is and ensuring that this functionality can be effectively communicated to system developers. Much as modelling methodologies transformed system and database design, we need to develop preservation modelling methodologies that can be used to formally represent preservation functions and processes. Moreover, most modelling work has attempted to replicate traditional processes. Formal modelling of preservation processes is the necessary first step in re-engineering them and developing tools for automatic management of digital entities. It would be desirable to define a generic preservation process/architecture so that similar processes can be applied across different domains.

5.2B: Automation of Processes: Current digital preservation processes require extensive human intervention for selection, validation, description, assigning unique identifiers, data management, migration, and delivery of desired content. The degree of human labour currently

required does not scale to the size or complexity of the digital content that needs to be preserved. Projects that develop processes and tools to support automated ingest, validation, metadata creation and data management are needed to make long-term preservation efficient and affordable. What are the particular capabilities required to automate appraisal, accession, description, arrangement, preservation and access? Are there additional archival processes that have not yet been implemented by the persistent archive community?

5.2C: Detecting Trustworthiness and Information Quality*:* Preservation processes often require transforming the original bitstream. Unless handled correctly, such transformations call into question the authenticity, quality and trustworthiness of the preserved entity. Authenticity can be protected in spite of data and system migrations by the quality of reconstructability of an information object. As far as static objects are concerned, measures like employing error-correcting codes or redundancy might help, but if we regard the algorithmic character of complex information objects which generate – or reconstruct – the visualisation the user perceives as the 'document' he/she is currently accessing, a more sophisticated approach is needed. A basic starting point is a definition of the attributes on which judgments about authenticity and quality are made. How is the quality of reconstruction of digital entities through migrations characterised as a function of the authenticity of the digital entity? What level of information loss is acceptable?

5.2D Scalability*:* Most digital preservation research to date has examined either large sets of homogeneous data or small collections of heterogeneous material. This raises a series of issues concerning the scalability of current models and methods, including the maximum archive size, the ingest rate, and the rate at which digital materials can be normalised or migrated. Can repositories be designed to handle collections with billions of digital entities in dozens of different formats? Is it possible to develop metrics to assess the scalability of preservation strategies and methods? For example, a crawl of five federal department Web sites in the United States retrieved 16.9 million digital entities, representing 233 different data types. Can the preservation of this collection be automated despite the large number of data types?

5.2E: Collection Completeness and Anomaly Detection: Most processes for validating the completeness and closure of collections are manual processes with limited ability to detect missing items, errors or

other anomalies. Can automated processes be developed as part of the accessioning process that would provide better detection of problems with collections? Is it possible to detect when collections are incomplete?

5.2F: Distributed and Grid Storage: New methods for distributed storage offer potential benefits to long-term preservation, such as increased redundancy and lower storage costs. Grid technologies, for example, provide support for methods that fragment and distribute digital entities, while reducing the risk that they are unnecessarily duplicated. What approaches are needed to build distributed systems such that digital holdings are distributed across geographically remote sites? How can archival processes be managed in distributed environments? How can the contents be reconstructed? What is the optimal size of a fragment unit? What impact does distributed storage have on ingest, documentation and delivery of digital entities? How do we track the state of distributed digital entities?

5.3 Preservation Systems and Technology

There are many opportunities to develop systems, tools and technologies to support digital preservation. Development of systems and technologies for digital preservation rests, in part, on producing formal models of preservation processes. This area also has the greatest potential for commercialisation.

5.3A: Formats of Digital Entities: Different formats require different kinds of strategic approaches to ensure that they can be accessed in the future. Problems with formats are exacerbated by the fact that archival collections, which need to be managed as a whole, generally contain entities in multiple formats; these formats have different rates of obsolescence. As well as producing evidence as to the properties that enable or put preservation at risk we need predictive measures to enable developers to assess the preservation impact of attributes of formats in advance of their completed development.

5.3B: Managing Complex and Dynamic Digital Entities: There has been little research to address how interrelationships between the components of compound documents might be maintained. How can complex and dynamic entities be authenticated and their integrity verified? How can dynamic entities be accessioned and managed in an archive? To what aspects of a dynamic document should metadata be attached and what metadata would be required?

5.3C: Automated Metadata Creation: Preservation metadata is an essential part of the information infrastructure necessary to support all the processes in digital preservation. To bring the preservation of digital objects closer to realisation, automatic or semi-automated creation and authoring of the metadata is a crucial issue. Based on existing technologies for extraction of low-level features and structural and administrative information, the challenge is the development of methods and approaches for creation of metadata supporting the understandability of digital objects. What tools can aid in the creation, authoring and management of metadata? At what time in archival processes should those tools be applied? There is also a need for harmonisation and integration of metadata approaches.

5.3D: Long-term Metadata Viability: The meaning of metadata itself changes over time, what we might describe as 'metadata drift'. For purposes of interpretation and authenticity, users will need access to the metadata schema used at the time the digital entity was created. There is a need for research into metadata schema and ontology evolution mapping to ensure that, over time, metadata and underlying ontologies do not lose their meaning. Tools are needed to track the provenance of metadata schema, for version control, and to allow users to navigate from current metadata schema and ontologies to those used when the digital entity was created. There is also a need to assess the value of metadata compared with the costs of extracting, creating and managing metadata that could lead to definitions of the minimum amount of metadata necessary for digital preservation.

5.3E: Multilingual Entities and Technology: Digital entities worthy of preservation come in all languages, but there has been little research on issues of multilingualism. Many archival collections include documents in multiple languages. Issues of multilingualism are not limited to the contents of collections. There is also a need for research on underlying metadata schema, descriptive terms, user documentation and interfaces.

5.3F: Acceptable Loss: We take as a given that it will not always be possible or economical to preserve all of the features and functionality of original digital entities, but we lack metrics for measuring what is acceptable loss. This is in part because we lack formal ways of representing functionality and behaviour. How can we measure what loss is acceptable? What tools could be developed to inform future users about the relationship between the original digital entity and what they

receive in response to a query? These questions are related to re-engineering appraisal and selection processes that will need to take into account that essential elements of digital entities need to be preserved.

5.3G: Repurposing: A new challenge on the action level consists in the re-purposing of archives. Reorganising archives and using their contents in unforeseen scenarios requires more than the usual access functions supporting search and presentation of found items. Whether future user communities can exploit stored assets might depend on the analysis tools they can use to find relevant information patterns in the collection, therefore adequate tools for information discovery are mandatory. Parallel research on discovery, search, retrieval and presentation can be applied to this problem.

5.4 Research Methodologies

Here we highlight some of the distinctive attributes of digital preservation research, recommend some general principles to guide the research process, and offer a view of the research cycle and the role of partnerships as considerations in defining the research agenda.

The research cycle for digital preservation needs to support pragmatic as well as theoretical work. The collaboration of key stakeholders is needed throughout the cycle, with each taking the lead when and as appropriate. Pathways from research to practice must be defined and utilised that build upon and presume stakeholder interaction. In the digital preservation domain, it is important to remember that practice informs research in a fashion similar to the ways in which research informs practice.

Examples of 'Good Clinical Practice' and 'Good Manufacturing Practice' that various domains have developed to ensure reliable results should inform the development of good research practice for digital preservation. There are aspects of research methodology that cut across domains, and these commonalities should be made overt. In defining good research practice for digital preservation, the resulting principles and guidelines should stress:

- the development of and adherence to standards for policies, procedures and practice;
- the need for interoperability as an essential requirement for the developments that result from the corpus of research; and,

- recognition of the coalescence around a set of community-endorsed frameworks, models and protocols.

The codification of good practice should enhance research results and enable the widest impact of funded research, while not being restrictive or inadvertently stifling innovation. The primary goal in developing good digital preservation research practice is to produce credible (internally valid), transferable (externally valid), dependable (reliable) and confirmable (objective) research results [26].

6.0 Additional Support Requirements

This section discusses additional needs that are not addressed directly by the research agenda.

6.1 Organisational, legal and policy issues

Some of the institutions responsible for maintaining cultural memory are becoming increasingly aware of the challenges associated with preserving digital objects and are starting to develop and implement an appropriate organisational and technical infrastructure. Unfortunately, however, most of the archives, libraries, museums, as well as content and service providers like private archives, broadcasters and companies, have not yet acted. The challenges they are faced with, and need support for, are:

- Building up expertise: There is a need to build up expertise concerning preservation of digital objects, especially in small, highly specialised companies. At the moment ad-hoc solutions for each specific environment predominate.

- Developing an organisational framework: Based on the introduced organisational methods for preserving physical objects in filing departments, libraries and archives, new methods reflecting the special requirements of born or converted digital objects need to be developed.

- Obtaining organisational buy-in: preservation requires cultural change on the part of information creators and this depends upon not only raising of awareness, but increasing the participation of creators in the process.

- Overcoming uncertain legal positions and policies: Legal issues could become one of the major obstacles to introducing long-term preservation. Only in the presence of clearly defined rules and policies, which mandate ownership of digital objects to the producer, will producers be willing to participate in the preservation process of their digital objects.

- Overcoming the heterogeneous and incompatible structure of archives: For example: Large, distributed organisations such as the broadcast industry have no common rules on how to describe the archiving of digital objects. As a result, archives within organisations exist where no information exchange and re-use can be realised. Overcoming this deficiency requires technologies which enable the integration of existing digital objects.

- Developing cost modelling tools: Ensuring organisational support for preservation and enabling longer-term planning for the revenue implications of engaging in preservation depends upon the availablity of cost modelling tools. These are currently lacking.

- Technology development: Preserving digital objects requires technology for preparing the objects for long-term preservation. As a consequence, not only does the question of availability of this technology arise, but also its applicability in small and medium-sized companies.

6.2 Technology Transfer and Commercialisation Scenarios

In contrast to many other research programmes, the primary objective of enquiry into preservation of digital content is practical rather than theoretical – to devise methods that come into practice so that resources are available for future knowledge generation. To be effective, each successful project must include specific measures to transform its results into forms that are satisfactory to its eventual users and, in some cases, transfer such 'deliverables' to organisations that plausibly commit to maintaining them for delivery to archiving enterprises or to the public. For instance, a project whose output is partly realised in software should plan for and report on its activities to ensure that this software survives beyond the existence of the project team.

It is important to recognise the profound distinction between prototype/pilot versions and 'industrial strength' versions suitable for integration into pre-existing environments with sufficient scaling, error handling, user education, and support infrastructure to satisfy different deployment environments. 'Industrial strength' software differs from prototype and pilot software less in its functional features than in its ability to scale from very small to very large data collections, and in having a support infrastructure for customer service.

Preservation technology can foster new markets. Opportunities lie in the realisation of new services, which can be offered by the digital content owners or by external vendors. Such a model can also be extended to include personal and public use. One possible commercial perspective could lie in a market for content providers offering storage, preservation of, and access to digital objects, and a market of service providers offering value-added services.

We expect technology transfer to cost a significant fraction of the resources of most projects funded under this initiative – from 25% to 50%. We recommend that research contract applications be required to propose how their teams will ensure that their work is embodied in durable solutions. We further recommend that project progress reports be required to describe accomplishment of technology transfer.

Conclusions

Future development and the sustainability of the information society require solutions to preservation challenges. Investing now will save money in the long term, for government, industry and the citizen, both by ensuring the preservation of intellectual capital and by obviating the need for each organisation to discover solutions for themselves. The outcomes of much of this research agenda not only provide a framework to improve preservation, but will underpin the development of new commercial opportunities in the form of both services and products. These benefits can only happen if the research is well-founded and conducted with a view to applicability and dissemination, both of which points are stressed in this report.

Selected References

[1] Abrams, S.L. and Seaman, D. (2003), 'Towards a Global Format Registry', IFLA2003 (Berlin), http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

[2] Beagrie, N. and Jones, M. (2001), *Preservation Management of Digital Materials: A Handbook*, London.

[3] Branschofsky, M. and Chudnov, D., (2002), *Dspace: Durable Digital Documents long term-preservation*. Proceedings of the Second ACM/IEEE-CS joint conference on Digital Libraries, Portland, Oregon, USA.

[4] Consultative Committee for Space Data Systems (January 2002), CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS)*. Blue Book. Issue 1, 1-1. http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

[5] CEDARS Project. *Metadata For Digital Preservation The Cedars Project Outline Specification Draft For Public Consultation*, http://www.leeds.ac.uk/cedars/metadata.html

[6] Friedlander, A. (2002), 'The National Digital Information Infrastructure Preservation Program,' *D-Lib Magazine*, 8.4, April 2002, http://www.dlib.org/dlib/april02/friedlander/04friedlander.html

[7] Hunolt, G. and Booth, A. (9/2001), *ESDIS Data Center Best Practices And Benchmark Report*, Science Operations Office, Earth Science Data and Information Systems Project, Goddard Space Flight Center, NASA Contract NAS5-00154.

[8] Hennessy, E. (2000), *COTS Supportability Across the Program Lifecycle: A Look At The COTS Evolution in the DoD Market*.

[9] Hodge, G.M. and Carroll, B.C. (1999), 'Digital Electronic Archiving: The State of the Art and the State of the Practice.' A Report Sponsored by the International Council for Scientific and Technical Information, Information Policy Committee and CENDI, 26 April 1999, http://www.icsti.org/99ga/digarch99_TOCP.pdf

[10] Holdsworth, D. and Wheatley, P. (2001), Emulation, preservation, and abstraction. CAMiLEON Project, http://129.11.152.25/CAMiLEON//dh/ep5.html

[11] It's About Time: Research Challenges in Digital Archiving and Long-term Preservation. Final Report on a Workshop on Research Challenges in Digital Archiving, 12-13 April 2002. (2002), Washington, DC, Library of Congress.

[12] *InterPARES Preservation Task Force Final Report*; Draft for Comment (October 2001), http://www.interpares.org/reports.htm

[13] Kampffmeyer, U., CeBIT 2002 Review, http://www.doculine.com/news/2002/0402/cebit-2002-review.shtml

[14] Keller, M.A., Reich, V.A. and Herkovic, A.C. (2003), 'What is a library anymore, anyway?', *First Monday*, 8.5, May 2003, http://firstmonday.org/issues/issue8_5/keller/index.html

[15] King, A., *Commercial Off-the-Shelf Software – Benefits and Burdens*, http://www.mitre.org/pubs/edge_perspectives/march_01/ep_king.htm

[16] *Kodak and Lockheed Martin Partner to Advanced Imaging Technology for Government and Commercial Applications. Digital Preservation Effort Seek to Build On Success of U. S. Census Program*, (2001), http://www.kodak.com/US/en/corp/pressReleases/pr20010314-01.shtml

[17] Library of Congress (2002), *Plan for the National Digital Information Infrastructure and Preservation Program*, Washington D.C. http://www.digitalpreservation.gov/index.php?nav=3&subnav=1

[18] Lorie, R.A. (2002), *A Methodology and System for Preserving Digital Data.* Proceedings of the Second ACM/IEEE-CS joint conference on Digital Libraries, Portland, Oregon, USA.

[19] *Metadata for long term-preservation.* (July 2000), http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm

[20] Moore, R., Baru, C., Rajasekar, A. et al. (2000), Collection-Based persistent Digital Archives. *D-Lib Magazine*, March and April 2000. http://www.dlib.org/dlib/march00/moore/03moore-pt1.html and http://ww.dlib.org/dlib/april00/moore/04moore-pt2.html

[21] Payette, S. and Staples, T. (2002), 'The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services', in M. Agosti and C. Thanos (eds.), *ECDL 2002*, LNCS 2458, 406-421.

[22] *Preservation Metadata and the OAIS Information Model.* A Metadata Framework to Support the Preservation of Digital Objects.A Report by the OCLC/RLG Working Group on Preservation Metadata, June 2002, http://www.oclc.org/research/pmwg/pm_framework.pdf

[23] *Preservation Metadata for Digital Collections*. October 1999, http://www.nla.gov.au/preserve/pmeta.html

[24] President's Information Technology Advisory Committee (PITAC). (2001), *Digital Libraries: Universal Access to Human Knowledge*, http://www.ccic.gov/pubs/pitac/pitac-dl-9feb01.pdf

[25] *PRESTO – Preservation Techologies for European Broadcast Activities. Existing and Emerging Technologies*, May 2001, http://presto.joanneum.ac.at/Public/D31.pdf

[26] Reason, P. and Rowan, J. (eds.), (1981), *Human Inquiry* , New York: John Wiley, 237-242.

[27] *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*,

(2003), Arlington, Virginia, USA: National Science Foundation, (January), http://www.cise.nsf.gov/evnt/reports/toc.htm

[28] Rothenberg, J. (1995), Ensuring the Longevity of Digital Information. *Scientific American*, 272(1):24-9.

[29] Schade, D. (2002), 'The Virtual Observatory: The Future of Data and Information Management in Astrophysics', 2002 CODATA Conference in Montreal.

[30] Sharp, R. (2002), 'The Ephemeral will Endure: The Future of Conceptual Art and Digital Preservation: An Interview with Jon Ippolito of the Variable Media Initiative at the Guggenheim Museum, New York (August 2002)', *DigiCULT.info*, 2 (October), http://www.digicult.info/downloads/digicult_info2.pdf

[31] Steenbakkers, J. (2000), in *The NEDLIB Guidelines - Setting up a Deposit System for Electronic Publications*, NEDLIB Report Series, 5, Koninklijke Bibliotheek.

[32] *The DigiCULT Report*. (2002), Technological landscapes for tomorrow's cultural economy. Unlocking the value of cultural heritage, http://www.digicult.info/pages/report2002/dc_fullreport_230602_screen.pdf

[33] Thibodeau, K. (2002), *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*. In: The State of Digital Preservation: An International Perspective.Conference Papers and Documentary Abstracts, http://www.clir.org/pubs/reports/pub107/thibodeau.html

[34] *Trusted Digital Repositories: Attributes and Responsibilities*. An RLG-OCLC Report, May 2002, http://www.rlg.org/longterm/repositiories.pdf

[35] Webb, C. (2002), *Digital Preservation – A Many-Layered Thing: Experience at the National Library of Australia*. In: The State of Digital Preservation: An International Perspective. Conference Papers and Documentary Abstracts, http://www.clir.org/pubs/reports/pub107/webb.html

Appendix A

Working Group Membership and Process

Margaret Hedstrom, University of Michigan (Co-Chair)
Seamus Ross, HATII, University of Glasgow (Co-Chair)

Kevin Ashley, University of London Computing Centre
Birte Christensen-Dalsgaard, Statsbiblioteket (Denmark)
Wendy Duff, University of Toronto
Henry Gladney, HMG Consulting
Claude Huc, French Space Agency
Anne R. Kenney, Cornell University
Reagan Moore, San Diego Supercomputer Center
Erich Neuhold, Fraunhofer (Darmstadt)

The Working Group met three times: twice in Washington and once in Paris. The group developed a matrix of current digital preservation projects and research initiatives at its first meeting in April 2002. The group also discussed the gaps in current research and developed a work schedule. Three month later, in Paris, the Group focused on developing a conceptual framework, reviewing the work done to date, drafting an outline for the final report and determining its content, as well as finalising a work plan for writing the report. A final meeting was held in Washington in November 2002 to review and finalise the report, and develop an action plan.

Acknowledgements of other contributors

Hans Hofman, National Archives of the Netherlands
Jen Engleson Lee, University of Michigan
Nancy McGovern, Cornell University
Ulrich Thiel, Andrea Dirsch-Weigand, Wolfgang Putz, ISPI Fraunhofer