National Archives of Sweden

Digital Preservation in Archives: Overview of Current Research and Practices

January 2004 - February 2005

Raivo Ruusalepp

Table of Contents

1. Introduction

"The last decade and a half has produced more records than any previous similar period of human activity. The fact, that the majority of these records is less reliable, retrievable or accessible than ever before, is one of the ironies of the modern information age."[1] Idiosyncratic software systems generate, manage and store digital data using proprietary technologies and media that are not developed to prevent manipulation and that are subject to obsolescence. Therefore, long-term preservation of digital information is plagued by storage media short lifespan, obsolete hardware and software, proprietary file formats, defunct technologies and web sites. Indeed, the majority of computer products and services on the market today did not exist five years ago.[2] The essence of the "problem" with digital preservation is the lack of proven methods to ensure that the digital information will continue to exist, that we will be able to access this information using the available technology tools, or that any accessible information is authentic and reliable.

The last twenty five years have seen vigorous research into issues of digital preservation: technological obsolescence, storage media fragility, the manipulability of electronic systems that challenge our capacity of guaranteeing the long-term preservation, and the authenticity of electronic records. The formidable body of literature that has accumulated on the topic is divided between (sometimes-conflicting) suggestions:

- to study and understand the technological context of electronic records in each individual case;
- that the preservation of authenticity of records over time should be based on requirements and procedures that are independent of specific technological contexts;
- to use platform-independent and open, standardised file formats for preservation;
- to emulate the original technological environment on future technology platforms;
- to migrate the records into formats suitable for access on current platforms;
- preserve the contents of digital records on paper or microfilm;
- etc.

Practical implementations of these (theoretical) suggestions by archives are twofold: on the level of policies and strategies; and *ad hoc* practical solutions where time for planning has been short. Rarely have the two been joined into a whole, as the following report testifies, with insufficient funding often cited as the main cause. Resources devoted to preservation of electronic records within archival institutions have not been commensurate with the task. Policy and strategy are fine, but without implementation they are not worth very much.[3]

Implementation is arguably the greatest unresolved issue of digital preservation, and the most difficult to deliver, because it involves major resources, compliant organisations, dedicated management and appropriately skilled staff. Archivists have sometimes felt that they are expected to provide answers and solutions to problems that outstretch their capabilities.

1.1 Archives as a stakeholder

Naturally, archives are but one interest group involved in finding solutions to the questions of long-term digital preservation – librarians, science communities, cultural heritage, businesses

[1] L. Duranti, *'From Here to Eternity': Concepts and Principles for the Management of Electronic Records* (1999), p. 1
[2] S.S. Chen, *The Paradox of Digital Preservation* (2001), p. 2
[3] G. O'Shea, *Research issues in Australian approaches to policy development* (1997), p. 253

and government agencies, technology industry, medical institutions and many others are seeking to overcome the same problems with technology obsolescence and extending usability of digital data that depends on this technology. Many of these stakeholders have their specific requirements and correspondingly a specific angle to the digital preservation problem. For many years, different interest groups have worked on their theories and solutions separately, but in recent years, different communities have started to co-operate again and the results of this are beginning to emerge. The specific needs of different stakeholders have also become clearer and better defined through this co-operation, and this includes archives.

For archives, public or private, the provision of their "traditional" services of providing long-term access to authentic and usable records, has become considerably more complicated through the use of digital technology. And in developing their "new", digital services, the archives have to bear in mind their own stakeholders who have different interests. For example:

- The users of archives
  Who have ever rising expectations to the access to resources in archives: access should be electronic, rapid, precise and preferably on-line.

- The records' creators
  Who have invested into technical infrastructure for creating records electronically and who clearly want that functionality of their records maintained also in the archives and for the long term.

- The society in general
  Digital records form part of the society's memory and need to be kept for future generations as a proof and sign of our time and generation.

- The heritage industry
  In general public perception archives are often grouped together with 'heritage organisations' or 'memory organisations', alongside with libraries, museums and other institutions that "do preservation". Many archives are making various collections accessible to the general public on-line through their digitisation programmes. Digitised material, even if created only for providing access to digital copies of original materials, tends to increase the need for digital preservation in the archives.

- Technology
  Technology that keeps developing and changing and making the archivists' task of preserving access to digital resources easier and more difficult at the same time. First serious attempts to air the archivists' concerns with the technology development to the ICT industry were made by the DLM-Forum[4] and were met with a supportive response.[5]

- Fundmakers
  Archivists and records managers need to constantly convince their paymasters that the solutions they are suggesting offer a good return on investment. Archives have occasionally suffered a temporary "loss of voice" because the answers to all the problems with managing and preserving digital records are not easy to find, particularly in a situation where the solutions sought by governments have to be clearly defined, cheap to implement, easy to use and reliable for long-term. These answers and solutions are only

[4] see: http://europa.eu.int/historical_archives/dlm_forum/doc/dlm-message-to-industry-en.pdf
[5] http://europa.eu.int/historical_archives/dlm_forum/doc/ictindustryresponse-en.pdf

being developed and, in the meantime, there is little to copy from and only a few who can be asked for guidance.

Digital environment has, perhaps more than ever before, prompted a need for archivists to come up with completely new solutions in a very short space of time, borrow theories and practices from other areas and collaborate with other disciplines that have not been traditionally close allies of archivists. This has launched an almost existential discussion of what it is the archives are doing but has also resulted in new approaches being developed, new funding opportunities, new methods of organisational co-operation, new ways of thinking. Archivists now understand better that in order to "sell" their needs and requirements for records processes, they need to develop convincing theoretical solutions and have to learn how to implement these into practice.

Digital preservation is a complex problem and solving it has to start already with the creation of a digital record: "the time between a object's "creation" and its preservation is shrinking rapidly for digital materials [and] preservation will need to be addressed increasingly at the time of acquisition or even creation of the digital resource".[6] In order for an archive to make informed strategic decisions for its digital preservation services, it needs to consider the records life-cycle as a whole, including the stages prior to preservation in archives and accessing and using the records in the archive. Preservation is one part of an integrated service that archives are providing to their users, and digital preservation is with growing significance in this integrated whole.

There is still no 100 per cent guaranteed solution for long-term digital preservation, yet. Nevertheless, an array of approaches and methods to reduce the risk of inaccessibility of records in the future, are available and the list is growing every year. Research into digital preservation issues has not stopped and many new, innovative solutions will continue to emerge, funding opportunities permitting.

1.2 Structure of the report

In December 2001, the Swedish Riksarkivet commissioned a report about the state of digital preservation research around the world as part of the wider archival investigation into improving archival services in Sweden. The purpose of the report was to provide an overview of what areas of digital preservation are currently being researched, what are the main outcomes of this research, what organisational models are being developed for both research and preservation itself and what of these achievements might be useful in the Swedish context. The report was to "take the pulse" of digital preservation research in order to help the Riksarkivet in defining its own requirements and deciding on future strategies for preserving digital records. The current text is an update of the original report presented to the Swedish National Archives in 2002.

The report takes a broad look at the digital preservation problem that begins with the creation of digital resources and ends with their usage in the archives. The report is structured (loosely) according to the life-cycle of a digital record, relating the different stages in the life-cycle of a record to the digital preservation issues from an archival perspective. The general approach to the digital preservation in this report is from an archives' point of view and addressing mainly the aspects of this area that have concerned archivists. The report includes descriptions of research in a selection of countries and by specific research projects.

[6] S. Granger, et al., *Cost elements of digital preservation* (2000)

The report was prepared and updated by Raivo Ruusalepp, consultant with the Estonian Business Archives, Ltd., between December 2001 and February 2005. I would like to thank Magnus Geber, Göran Kristiansson and Christina Olsson from the Swedish Riksarkivet for their practical help as well as comments on the text of the report.

Please address your comments and questions to Raivo Ruusalepp at: raivo@eba.ee.

2. Digital records in records management

Traditionally, the archives have had limited means of impacting the way records are managed prior to their transfer to the archives – the legal context and status of (national) archives is slightly different in every country. The widespread use of electronic records has made this a significant problem and archives in many countries have begun finding new ways to regulate, guide and audit their depositor organisations' management of electronic records.

2.1 Guidance and consultation of records creators

The ways and means for creating and managing digital records prior to their transfer to the archives for long-term preservation, are important from the archives' point of view because they have a direct impact on the quality of records that the archive will accession. In essence, archives are interested in exercising certain "quality control" over the life cycle of the material that they will have to preserve and provide long-term access to. The other purpose with getting involved in the earlier stages of the records life-cycle control is to distribute the responsibility for the digital preservation among all stakeholders who are involved with the same digital records. Archives alone cannot be responsible for the long-term preservation of digital records if these are created in unsuitable formats or if their condition has deteriorated through inappropriate handling during usage or storage.

Archives are participating more and more in regulating the complex synergy of technical and organisational issues that surround the creation of electronic records and their management throughout the active stages of life-cycle. Three levels can be identified in the control that archives are exercising over the earlier stages of the record life-cycle:

- Setting legal or recommended requirements
  In many countries the remit of archives does not formally reach further than setting requirements for the records transfer stage. Specific guidelines for transferring electronic records to the national archives have been issued in Denmark, Sweden, Norway, Germany (see also ch. 2.3).

  Following the publication of the Australian national Records Management Standard (AS 4390) in 1996, the National Archives of Australia has put much emphasis on the electronic recordkeeping systems that government agencies have been implementing in order to make sure that these systems contain features for managing records over time. The National Archives maintains an on-line manual on designing and implementing a recordkeeping system (DIRKS).[7] Guidelines for storing electronic records in organisations until their transfer to the national archives have also been published in Australia: *Protecting and handling magnetic media* (2002)[8], *Protecting and handling optical discs* (1999)[9]. General guidelines for managing electronic records have also been issued by the Swedish National Archives: *Riksarkivets föreskrifter och allmänna råd om upptagningar för automatisk databehandling (ADB-upptagningar)* (RA-FS 1994:2, 2003:2)[10]; by the National Archives of the Netherlands: *Regeling geordende en toegangelijke staat archiefbescheiden* (2002)[11]; by the National Archives in the UK:

[7] http://www.naa.gov.au/recordkeeping/dirks/dirksman/dirks.html, http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice51.html
[8] http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice5.html
[9] http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice6.html
[10] http://www.ra.se/ra/pdf/RAFS/RAFS%201994-2.pdf, http://www.ra.se/ra/pdf/RAFS/RA-FS%202003-02.pdf
[11] http://www.rijksarchiefinspectie.nl/pdf/regeling12.pdf

Guidelines for Management, Appraisal and Preservation of Electronic Records (1999)[12] and *Electronic Records Toolkits*[13]; etc.

- Issuing a functional requirements for electronic records management systems
  Compared to recommended best practice guidance and regulating the records transfer process, more control over the record life-cycle can be gained through setting requirements for systems that produce and manage records. Many national archives are either involved or in charge of developing functional requirements for electronic records management systems (ERMS).

  The National Archives of Canada was probably the first to issue clear requirements for electronic document and records management systems used in the government agencies: *Records/Document/Information Management (RDIM): Integrated Document Management System for the Government of Canada – Requirements* (1996).[14] The requirements were contextualised in more general treatments of records management in the digital age: *Electronic Work Environment (EWE) Vision*[15] and *Record Keeping in the Electronic Work Environment: Vision*.[16] Based on these texts, the official Canadian RDIMS – *Records, Document and Information Management System* – concept was compiled that by now comprises of over five different software products that the agencies are encouraged to use in order to guarantee good recordkeeping.[17]

  The UK National Archives (TNA) has issued already two versions of functional requirements for electronic records management systems: *Functional Requirements for Electronic Records Management Systems* (1999 and 2002).[18] The last revision includes, in addition to the requirements and a reference volume, a record management metadata standard and an implementation guide. The TNA requirements, together with a separate guidance text *Guidelines for Management, Appraisal and Preservation of Electronic Records* (1999) are, perhaps, the most thorough example of functional requirements for ERMS that are directed also to practical use by the agencies and organisations.

  Following a request of the DLM Forum, the European Commission's Interchange of Data between Administrations (IDA) programme tendered, in 1999, for a specification of functional requirements for the management of electronic records. Development of the functional requirements was carried out by Cornwell Affiliates plc, supported by a guiding team of experts from several countries, with international validation organisations from both the private and public sectors. The resulting set of model requirements for the management of electronic records was finished in 2001: *MoReq: Model Requirements for the Management of Electronic Records*.[19]
  The MoReq requirements contain a model of how file plans, files and records relate to each other. The model is thought to be applicable to both electronic and hybrid files that contain both electronic and paper records. MoReq also includes a metadata model.
  Originally produced in English, MoReq is now available in at least five different languages. The DLM Forum has set up a working group for further development and certification for the Model Requirements for Electronic Document and Records Management Systems (EDRMS).

[12] http://www.nationalarchives.gov.uk/electronicrecords/advice/guidelines.htm
[13] http://www.nationalarchives.gov.uk/electronicrecords/advice/
[14] http://www.archives.ca/06/docs/4rdims.pdf
[15] http://www.archives.ca/06/docs/2ewe.pdf
[16] http://www.archives.ca/06/docs/3rk.pdf
[17] http://www.rdims.gc.ca/index_e/RDIMS-overview.pdf
[18] http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/
[19] http://europa.eu.int/ISPO/ida/export/files/en/635.pdf

The National Archives in the United States have formally endorsed and given suggestions for improvements[20] for the standard developed by the U.S. Department of Defence: *Electronic Records Management Application (RMA) Design Criteria Standard*, DoD 5015.2-STD (1997, 2002).[21] After careful consideration, the National Archives have recommended the standard functional requirements for use by all government agencies.

Functional requirements can be used for certification of electronic records management software systems for compliance with the requirements. In the UK, the National Archives has taken on the role of ERMS certification, in the U.S. the Department of Defence runs a similar programme. The Canadian RDIMS "package" consists of software products already certified as compliant with the RDIM requirements. A form of systems certification is also in place in Denmark, although there are no formally established functional requirements for the ERM systems. Decree of the national archivist *Anmeldelse og godkendelse af elektroniske journaler og dokumenthåndteringssystemer* (2002) and a user guide to it,[22] make it mandatory for all government agencies to get their records management and recordkeeping systems approved by the National Archives within three months of their implementation or significant alteration. The National Archives has developed a form for describing the system that the agencies need to fill in and present to the Archives.

- Perhaps the most effective control over the earlier stages of a record life-cycle can be exercised through standardisation of records management processes. Standardising record management and -keeping practices in all government agencies is a very complex and laborious task and has been achieved only in a few smaller countries. Larger countries have developed only general models of records management. A standard or model that determines what functionalities a software package used for managing records must provide and how it should perform these functions, can be used to guarantee that electronic records are created and managed in a way that makes it possible to preserve authentic electronic records for long term in an archive. Full record life-cycle control and full consideration of archival requirements for preservation of electronic records has been the central part of the few examples of such standards.

  The Norwegian freedom of information act of 1969 made it mandatory for all government agencies to make their diary systems (registries of incoming and outgoing documents) publicly accessible. Over the years this requirement led to the automation of diary systems and a trend towards standardisation of records management processes associated with the diary systems. As a result, the first version of the NOARK standard (*NOrsk ARKivsystem*) was enacted in 1984. The main responsibility for developing the standard lies with the National Archives of Norway that has by now published four versions of the standard. NOARK-4 was published in 1999 and most of it has been translated into English.[23]
  NOARK sets requirements for records management systems':

  1) informational content – what records they must be able to capture and manage;
  2) data structure – relations between individual data elements;

[20] http://www.archives.gov/records_management/initiatives/dod_standard_5015_2.html
[21] http://jitc.fhu.disa.mil/recmgt/standards.htm
[22] http://www.sa.dk/sa/statamtkom/arklov/nyanmcirk.htm and http://www.sa.dk/sa/statamtkom/vejledninger/Anmvejljour.pdf
[23] http://www.riksarkivet.no/noark-4/Noark-eng.pdf

3) functionality – what the systems must be able to do;
4) in some cases, also to the user interface of systems and how the user must be able to interact with the system.

NOARK has a specialised module for transferring records to the National Archives according to a standard set by the Archives. The NOARK version 4.1 (2002) updated this module to include an XML-based transfer form for records and metadata.[24]

The Swiss federal administrative tradition relies on the same instrument as the German tradition — the registry plan (*Registraturplan* or *Aktenplan*) that relates particular activities with their respective functions. Such a registry plan reflects the hierarchy of functions and business processes of an agency as it performs its activities.
Since 1988 the Swiss Federal Archives has been involved in developing a core data model for automating the registry plan in a way that meets the archival requirements in the electronic environment. The project is called *Geschäftswervaltung* — GEVER; it is a joint project between the Federal Archives and the Federal Office for Information Technology.[25] It released its first records management standard in 1995. Further regulations for records management were released in 1998 and a revision of GEVER standard appeared in 1999; latest updates of the standard were issued in 2003.[26] The standard provides a framework of processes and function for records management systems. Aside records management, functions are provided for workflow management and its control mechanisms. Based on the 1999 version of the GEVER standard a new project GBL99 (*GEVER Basis Lösung*) was started with the aim of developing a practical implementation of the standard and a technical specification for systems and tools that are to be built after this standard model. In co-operation with software companies UNISYS and Fabasoft some electronic records management systems where adapted to meet the GBL99 requirements[27] and their data model was published as a general GBL standard in 2001.[28]

The State Archives of Basel-Stadt canton has also chosen the way of standardisation of records management processes. A project called PRISMA (*Produktivitätssteigerun dank Informationssystem im Archivbetrieb*) that ran between 1996 and 1999, developed an archival information system with functionality for transfer of records (including electronic records) from agencies to archives and a user-module for users of archives. Based on the PRISMA project a new project, ELGAR (*Elektronisches Geschäfts- und Aktenregistrierungssystem*), was initiated in 1999 with the aim of developing a model registry plan system.[29] This records registration and record keeping tool is based on various standardised models that have been developed internationally (incl. NOARK, GEVER, MoReq, DOMEA) and is offering a software-independent business process integrated record keeping standard. Contrary to the previous two models, the ELGAR implementation was started as a bottom-up exercise with the prototype software being tested in real agency working environment and requirements for functionality were developed from the results of the testing.

The decision in 1991 to share administrative functions of the German Government between Bonn and Berlin prompted an urgent need for a geographically distributed, yet consistent way of managing government administration electronically. The Joint Berlin-

[24] http://www.riksarkivet.no/noark-4/noark-4_1.pdf
[25] http://www.isb.admin.ch/internet/gever/index.html
[26] http://www.isb.admin.ch/internet/informatikstandards/standardindex/01431/index.html?lang=de
[27] http://www.fabasoft.com/html/news/news/2001-02-22-ch.htm
[28] http://www.isb.admin.ch/internet/informatikstandards/standardindex/00461/index.html?lang=de
[29] http://www.bs.ch/stabs/main_projekte-e.html

Bonn Information System (*Informationsverbund Berlin-Bonn*, IVBB[30]) became the umbrella for many pilot projects in the areas of electronic document management, electronic records management, workflow systems, and digital archiving. The Federal Archives has been actively involved in several of these projects, most notably in the DOMEA project[31] — *Dokumenten Management und Elektronische Archivierung im IT-gestützten Geschäftsgang* — which has become the focal point of research, regulation and requirements for electronic records issues. Led by the Co-ordinating and Advising Agency for Information Technology at the Ministry of the Interior (*Koordinierungs- und Beratungsstelle der Bundesregierung für Informationstechnik in der Bundesverwaltung*, KBSt) and with participation of the Federal Archives, the DOMEA project has developed a concept for the electronic, "less-paper" office[32] that follows the German *Registraturprinzip* and sets out its principles for handling digital documents. The concept is formulated as a set of functions and data models which has been, since 2002, used for certification of software products that comply with the standard.[33]

The Victorian Electronic Records Strategy began as a project in 1995 at the Public Record Office of Victoria (PROV) in Australia to examine issues related to the long term preservation of electronic records. The initial investigation produced a report called *Keeping Electronic Records Forever* (1996),[34] which included a number of recommendations. To implement these, in 1998 the Victorian Electronic Records Strategy project (VERS) was started, and out of this work the comprehensive *Standard for the Management of Electronic Records* (PROS 99/07) has been developed. Following the publication of the standard, work was continued to develop functional requirements for recordkeeping systems and to develop software to meet the requirements. The software was tested as part of the VERS@DOI project[35] and a *VERS Toolkit* was produced with guidance for the agencies. The second version of the VERS standard was released in 2003[36] and includes a number of "Advises" that help the implementation process of the standard as well as software development for both managing electronic records and archiving them.

Archives in many countries have actively been developing requirements for records management standards and -tools. Many of the standardisation processes initiated by archives have later become parts of official e-government strategies of many governments gaining thus wider context and importance.

2.2 Metadata used in records management

Metadata created together with and for the records forms the basis of records authenticity control and choosing the preservation methods for them in the archives. The participation of archival institutions in setting requirements for records management metadata is, essentially, another form of records quality control – only electronic records with sufficient metadata can be preserved for long term. Archives are also interested in compliance between the records management and archival metadata standards.

[30] http://www.kbst.bund.de/Anlage302891/Informationsverbund+Berlin-Bonn+-+English+Version.pdf
[31] http://www.kbst.bund.de/dokumente/Publikation/,-300364/dok.htm
[32] http://www.kbst.bund.de/Anlage300422/Band+45+komplett+%28347+kB%29.pdf
[33] http://www.kbst.bund.de/Anlage300444/Band+53+komplett+%281.2+MB%29.pdf
[34] http://www.prov.vic.gov.au/vers/kerf.htm
[35] http://www.prov.vic.gov.au/vers/projects/versdoi.htm
[36] http://www.prov.vic.gov.au/vers/standard/default.htm

The International Standards Organisation (ISO) Technical Committee ISO/TC 46, Information and Documentation is about to approve a general technical report for records management metadata ISO/PDTR 23081 *Information and documentation - Records Management Processes - Metadata for Records - Principles*. The technical report does not contain metadata elements or schemas, but offers good practice guidelines for creating and maintaining metadata.

One of the pioneers among archives developing metadata schemas for records management was the National Archives of Australia, together with the provincial archives of the country. Co-operating with the *Records Continuum Research Group* at the Monash University, the National Archives released its *Recordkeeping Metadata Standard for Commonwealth Agencies* (1999)[37] as a common descriptive standard across government agencies. Compliance with the Recordkeeping Metadata Standard is intended to help agencies to identify, authenticate, describe and manage their electronic records in a systematic and consistent way to meet business, accountability and archival requirements.
As part of the *Victorian Electronic Records Strategy* (VERS) project PROV have developed a *VERS Metadata Scheme* (2003)[38] – an elaboration of the national recordkeeping metadata standard with added metadata for digital preservation. The New South Wales Record Office *NSW Recordkeeping Metadata Standard* (2001)[39] is an adaptation of the standard of the National Archives for use in provincial administration.

Led by the National Archives, the Canadian Information Management Forum developed a position paper *Approach to the Description and Classification of Government Records* (1999),[40] that presented main requirements for describing electronic records within the RDIM systems. Based on this document, a metadata standard was published a few years later – *Record Keeping Metadata Requirements for the Government of Canada* (2001).[41]

The Swiss GBL99 standard includes a section on records management metadata standard that the software systems complying with the GEVER model must support: *GEVER Basislösung Metadaten zu den Objekten Ordnungssystem, Dossier und Dokument* (2001).[42]

As part of the functional requirements for electronic records management systems the National Archives in the UK has developed a metadata schema *Requirements for Electronic Records Management Systems, part 2: Metadata Standard* (2002).[43] The new version of the standard (2004)[44] includes sections on digital preservation.

Recordkeeping and records management metadata schemas have been cloned from the existing ones and developed anew by many other institutions in different countries. Some examples include: *Minnesota Recordkeeping Metadata Standard* (IRM 20);[45] United Nations *ARMS Standard on Recordkeeping Metadata* (2003);[46] *South Australian Recordkeeping Metadata Standard* (SARKM) (2004);[47] etc.

---

[37] http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm
[38] http://www.prov.vic.gov.au/vers/standards/pros9907vers2/pdf/99-7-2_Std_ver2-0.pdf
[39] http://www.records.nsw.gov.au/publicsector/erk/metadata/metadata-std/NRKMStitle.htm
[40] http://www.imforumgi.gc.ca/new_docs/draft_e.html
[41] http://www.imforumgi.gc.ca/products/meta/metadata3_e.pdf
[42] http://www.isb.admin.ch/internet/informatikstandards/standardindex/00581/index.html?lang=de
[43] http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/pdf/metadatafinal.pdf
[44] http://www.govtalk.gov.uk/documents/Records_management_metadata_standard_2002.pdf
[45] http://www.mnhs.org/preserve/records/metamrms.pdf
[46] http://www.un.org/Depts/archives/ARMS_MetadataStandard.pdf
[47] http://www.archives.sa.gov.au/files/management_standard_metadata.pdf

Records management metadata cannot exist in isolation – within a single agency or records management system alone – they must be exchangeable, interoperable and be preserved for long term across several technical platforms. Hence, most of the metadata schemas developed by national archives are either linked to or integrated into models of records management systems (e.g., DOMEA, GEVER, NOARK, etc.); have become part of the government-wide information exchange systems (e.g., Government Information Locator Systems in the United States and Australia, RUPA network in Italy, etc.); or are tightly integrated with the national information exchange standards (e.g., e-GIF in the UK).

The specific interest on national archives at that is the compliance and interoperability of records management metadata with the archival description – a vital component of the transfer of electronic records into the archive is the (automatic) transfer of metadata records from agency to archive. The better the compliance between the metadata schemas, the less processing the transfer will require and consequently the cheaper it is. The participation of archives in the records management metadata standardisation is, thus, driven by the practical need to ensure comprehensive and even description of records throughout their life-cycle.

2.3 Digital signatures as records authentication method

Authenticity of digital records and its preservation is a subject of continuing discussion, but the authenticity criteria have not been conclusively defined yet. In records management practice, digital signatures and encryption techniques are widely used as methods for shorter term proof of record's authenticity. The use of digital signatures has in most countries been regulated with legal acts, but these do not, as a rule, cover the preservation of digital signatures in archives.

Digital signature technologies both offer solutions and raise new problems. Solutions include possibilities to identify the author of the record and check whether the content of the record has been changed. Combining digital signatures with time-stamps offers additional authenticity guarantees for records. Main problem has been the short "lifespan" of digital signatures (or their issuing certificates) and dependency on particular, often proprietary software solutions. Suggestions for preserving the digital signatures for longer term focus on preserving the public keys and their certificates together with the records, but national archives have, so far quite unanimously, been against preserving the functionality of digital signatures in the archive. This means that although digitally signed records may be accepted to the archive, but the possibility of authenticating the digital signatures (or their issuing certificates) will not be provided by the archive. It is a common requirement that all records transferred to the national archives must be free of any kind of encryption or specialised encoding, yet still authentic. The presumption, thus, is that the electronic records are authentic at the time of their acquisition and the archive will be responsible for their continuing authenticity throughout preservation in the archive.

Directives of the European Commission regarding the use of digital signatures (1997 and 1999) prompted the member states of the European Union to change their legislation to allow for use of digital signatures and create the public key infrastructure for certification of digital signatures. First of the digital signature laws were enacted in Germany and Italy (1997) and the national archives of these countries were the first to react to new requirements. The German Federal Archive requires decryption and "opening" of digitally signed records before their transfer to the archive, the records must be readable and useable without requiring any specialised software or keys, and the archive will not preserve the functionality of digital

signatures.[48] National archives in other countries, where new legal acts were passed mainly in 2000 and 2001, are sharing this approach. For example in Finland, where the certificates of digital signatures were, according to the law, deemed for permanent preservation, the National Archives contested this rule since the National Archives is the only agency apprising records for permanent preservation. The retention period of certificates has now been shortened to 10 years.[49]

The Canadian National Archives has issued *Guidelines For Records Created Under a Public Key Infrastructure Using Encryption And Digital Signatures* (2001)[50] with a requirement that all records be decrypted and made freely readable prior to their transfer to the archives. The National Archives will not provide digital signature verification services after the records have been transferred.

The National Archives of United States (NARA) has produced guidance material for agencies *Records Management Guidance for Agencies Implementing Electronic Signature Technologies* (2000).[51] The guidance document includes recommendations for implementing the digital signature technology, but at the same time sets a requirement that for records with a permanent retention must have name of the signer and time of signing recorded also in human-readable text (e.g., in metadata). Otherwise NARA will not accept the record for acquisition. A further guidance text advises agencies in implementing the PKI-infrastructure: *Records Management Guidance for PKI-Unique Administrative Records* (2003).[52]

One of the hindrances to the widespread use of digital signature technology in records management has been the lack of methods for long-term preservation of authentic digitally signed records. Archives, feeling under pressure from new arriving technology, have refused to take on new functions (preservation and verification of certificates of digital signatures, essentially acting as certification authority for the duration of the records retention period). Instead, archives have offered a combined solution where during the active stages of the records life-cycle technological means are used to prove its authenticity (e.g., digital signatures) and when the records reach the archive, organisational methods will be used to preserve the authenticity of records. Using the so-called audit trail of authenticity where all actions performed on a record are recorded does, nevertheless, not satisfy all groups of archives' users. First court cases where non-verifiable digital signatures have not been accepted as trustworthy proof, have prompted the technology providers to seek new solutions. The European Electronic Signature Standardization Initiative (EESSI) has offered on new, "preservation-proof" digital signature formats[53] and the concept of *Trusted Archival Services* (TAS),[54] that should guarantee certificates of digital signatures that can be preserved forever. No comments from archivists are available to these proposals as yet.

2.4 Appraisal of electronic records

Appraisal of records derived from electronic environments does not differ substantially from the appraisal of traditional records, but some subtle differences remain. In order to be able to preserve electronic records for long term, the archive must appraise the technical condition, context and authenticity of records, alongside their value. The evaluation of the technical

[48] M. Wettengel, A. Engel, *Disposition and archiving of authentic electronic records in the Information Network Berlin-Bonn* (2000), p. 107
[49] R. Pohjola, *Implications of electronic signatures – the situation in Finland* (2002)
[50] http://www.archives.ca/06/0618_e.html
[51] http://www.archives.gov/records_management/pdf/electronic_signiture_technology.pdf
[52] http://www.archives.gov/records_management/pdf/final_pki_guidance.pdf
[53] ETSI TS 101733: *Electronic Signature Formats* and ETSI TS 101903: *XML Advanced E-Signatures*
[54] http://www.law.kuleuven.ac.be/icri/publications/91TAS-Report.pdf

condition of an electronic record and finding out whether it has been altered or tampered with since its creation, generally require new skills from archivists and new approaches to the appraisal process.

A summary of recent appraisal-related, mostly theoretical, research has been compiled by the InterPARES project,[55] whose task force has also published a study on the impact of digital technology on appraisal.[56] One of the European Commission's DLM-Forum conferences was dedicated to the appraisal of electronic records.[57] National archives in several countries have issued guidelines and recommendations for new methods in appraisal.[58]

2.5 Conclusion

Most national archives have only a few digital records in their collections since the bigger "wave" of electronic office documents and records is still held by the agencies. And this is also the main drive for the archives to issue guidelines and regulate the creation, management and storage of electronic records in the agencies. Archives have initiated research projects for this purpose and are increasingly attempting to tie their own initiatives and projects with larger e-government developments undertaken by governments. The accomplishments in this process testify both the usefulness and success of the archival projects.

[55] http://archivi.beniculturali.it/Divisione_V/convenzioni/censimento/appraisal.html
[56] http://www.gseis.ucla.edu/us-interpares/pdf/AppraisalTaskForceReport.pdf
[57] http://www.narc.fi/dlm/
[58] http://www.nationalarchives.gov.uk/electronicrecords/advice/pdf/procedures4.pdf; http://www.archives.govt.nz/continuum/dls/pdfs/s1-standard-appraisal.pdf

3. From agency to archive

The transfer of electronic records from agency to archive and transformation of records to a state that is fit for long-term preservation in archives has both theoretical and practical implications. Thus far, the theory is more advanced than the practice. The latter derives largely from the methods developed in archives in the 1970's and 1980's and is often causing losses in functionality of records when they are archived. The archival theory does and probably should not develop as rapidly as the technology that is used for creating electronic records. However, the archival approach should still offer a method for capturing the necessary records from the wide spectrum of digital information types that the agencies are creating. The archival methods currently in use have been described as conservative, backward and destructive, because they are mostly based on fixing or freezing an electronic record to a static state or format that can be preserved for longer term. Some new ideas for capturing records for archiving have emerged from research projects recently.

3.1 What to capture?

The question "What is a record?" has been one of the fundamental issues in debates and discussions of digital preservation over the past fifteen to twenty years. Attempts to arrive at general definition of an electronic record made in the last decade by several research projects (e.g., Pittsburgh, UBC, VERS, partly also InterPARES I) have by now replaced by somewhat more constructive approach to define these properties and functionalities of electronic records that need to be preserved in a given context. An electronic record can be perceived to be a package or a set of technical properties and possibilities, and of administrative context where some of these technical possibilities were put to use. The purpose of digital preservation is, therefore, maintenance of the technical possibilities used in a given administrative context and their usability through time. Some of the technical possibilities of the electronic record, that were not used in the given administrative context can be ignored, on the condition that this does not jeopardise the authenticity of the record. For example, if a record deemed for long-term preservation is contained in a database-based web-site solution, it may be sufficient to archive only the web document and the database engine, the pop-up advertisements, news bullets, etc. visible on the web page may be left out; when an MS Word document is converted into PDF format; etc.

The InterPARES project, second phase of which began in 2002, is working on theoretical and methodological issues that are fundamental to preserving electronic records for the long term. Based on the research, model strategies and standards are formulated to help with the digital preservation process. The current project is divided into four strategic areas, two of which are closely related to significant properties of electronic records and their requirements (Authenticity Task Force, Appraisal Task Force). In their work, these working groups are tackling questions like: what are the common characteristics of all electronic records? what criteria are used to categorise electronic records? what characteristics of electronic records are vital to preserving its authenticity? At what stage in the life-cycle should appraisal happen and should it be done more than once? etc.
Studying the properties and characteristics of documents and records, the InterPARES Authenticity Task Force came to a conclusion that because of the complexity of electronic records and record keeping, it is both difficult and problematic for those researching or managing electronic records to identify a single, appropriate unit of analysis. Diplomatics approaches the issue from the perspective of the individual record; archival science, from that of the record aggregate; and systems analysis, from that of the automated information or recordkeeping system. Each of these perspectives contributes to both understanding the nature of the record and its long-term preservation. What is also required, however, is an overall

systems approach that takes into account the total record-keeping environment, that is, the sum of all of the identified contexts.[59] It is hoped that with the remaining two years, the project will work on developing such an overall approach.

A slightly different approach has been taken by the Victoria Public Record Office in Australia that has published its *Victoria Electronic Records Strategy* (VERS) standard (version 2, 2003).[60] The standard includes requirements that every electronic record must meet in order to enable the preservation of integrity of record's content, context and structure: functional requirements for recordkeeping systems, required metadata elements for records and their aggregates and the encapsulation format for records.
The theoretical framework that was set out in the PROV 1996 study report advocated the creation of a 'static record' that was inviolable, satisfied evidentiary requirements of court and government, and that records should be captured at the time of creation. The overall approach taken was data driven, rather than systems oriented, which led to the recommendation for a long term electronic record format. The recommended standard record format prescribed that data structure should encapsulate the documents, the context, and authentication in a single object. One of the two main developments after the 1999 VERS *Standard for the Management of Electronic Records* (PROS 99/07) was the definition of a long-term preservation format for electronic records. The format is known as VEO – VERS Encapsulated Object. The VEO encapsulates in an XML wrapper the binary (either PDF, TIFF) or ASCII format of the record and the capsule is digitally signed.[61] The VEO is designed to exist independently of the system used for creating or managing the record, can be preserved for long term (PROV considers long-term to be at least 100 years) and has to be equipped with necessary metadata from its creation. The planned digital preservation strategy for VEO's will be migration.

New Zealand offers an example of a guideline issued by the national archives regarding identification of records in complex record keeping systems.[62]

These two examples are only a fraction of the attempts to define the electronic record, its properties and an ideal format for its preservation. Both projects described here have admitted, that not all issues have been solved yet and the work is continuing. In the new VERS standard the VEO concept has been updated in several technical aspects that were not there when the first version of the standard came out (e.g., the mandatory digital signatures in the VEO). Standards similar to VERS – NOARK and GEVER, for example – that are more systems-centred, yet still define file formats for records management and archiving, are also in constant development and change with new versions having to encompass new trends in the developing technology.

It should be concluded that there is, as yet, no description of the electronic record, neither in theory nor working in practice, that would "free" the archivist from making choices based on the technical condition and properties of the record. In other words, the possibilities an archive has for preserving electronic records are still largely determining the choices what records can be preserved and what not. Crucial role in this choice is played by the significant properties of electronic records.

3.2 Significant properties of electronic records

[59] InterPARES, *Authenticity Task Force Report* (2002), p. 32
[60] *Management of Electronic Records*, PROS 99/007 (Version 2)
[61] http://www.prov.vic.gov.au/vers/standards/pros9907vers2/spec_03/default.htm and http://www.prov.vic.gov.au/vers/published/vers.dtd
[62] *What is a Corporate Record?* (2003)

Preserving the electronic records as bit-streams only, with no regard to their content, is not sufficient for the archives nor for their users. The level at which an archive can offer (future) access to electronic records is dependent on the type of the record, its properties and how well the archive can preserve the usability of these properties. The archive can only preserve the record with properties that it had at the time of transfer to the archive – the archival preservation does and should not augment the record or its properties.

The concept of significant properties (also essential characteristics, essential attributes) has so far been treated mostly from the technical aspects of a record – they have been understood as properties of the file, its format and the software its use requires. Through the efforts of the CEDARDS project (UK) and others, the concept of significant properties has been extended to include properties of the record as a conceptual whole, rather than just a file saved onto a storage media.

The Cedars project (*CURL Exemplars in Digital ARchiveS*)[63] began in 1998 as a collaboration between three university libraries – Leeds, Oxford and Cambridge. The main objective of the project was to address strategic, methodological and practical issues and provide guidance in best practice for digital preservation. Running altogether for four years, the project produced a number of guidance documents that summarise the main areas of research of the project: intellectual property rights, preservation metadata, collection management, digital preservation strategies and the Cedars distributed archive prototype.[64] Of the three main participants of the Cedars project, the University of Leeds Library is now involved with the CAMiLEON project.

Preservation of an electronic record always includes the identification of an abstract model of the record data that preserves all the necessary properties of the record data, i.e. the significant properties of the record. This abstract model Cedars calls the 'underlying abstract form'. In many cases, the choice of underlying abstract form is obvious, but sometimes this is not so, and several choices present themselves. Key to making the choice of underlying abstract form is the identification of the resource's significant properties. Each potential choice is then weighed against its ability to preserve these significant properties. For example, under UNIX the case of file names is a significant property, and should be preserved, but it need not be preserved in a Windows system; to confuse the names in a UNIX file system has the potential to lose data by overwriting files when copying the file tree; etc.[65] One way to perform this assessment is to imagine that the original hardware is still available, and to ask if one could — in principle — recreate a working copy of the preserved object from the data as represented in the putative abstract form.[66]

Different classes of digital objects will have different types of significant properties, for a complex digital object a lot of effort may go into deducing significant properties. The view of what is significant can also be subjective, but while knowing what we are preserving, it is important to note what we are not preserving or are consciously leaving aside. It is good practice to state any assumptions, and think of what is significant to us now — omissions may easily occur because they are the norm to us at the moment. The expectation is that the analysis of what is significant can often be carried out for a whole class of objects (e.g., MS Word files), and the preservation of a newly arrived digital object in such a format merely re-uses the analysis.

[63] http://www.leeds.ac.uk/cedars/
[64] http://www.leeds.ac.uk/cedars/pubconf/pubconf.html
[65] http://www.leeds.ac.uk/cedars/testSitePhase2/sigProp1.html
[66] http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html, ch. 13

The concept of significant properties of a record has more recently been enhanced to include the context of creation and use of a digital object that both also add significant, worthy of preservation properties (e.g., authenticity, provenance).[67] It is to be hoped that this new approach gives rise to new treatments of the archives ingest process and decisions it precludes.

3.3 Transfer of electronic records and its legal regulation

In order to make it easier for national archives to control the transfer of electronic records from agencies and the processing of records, many countries have augmented their legislation. The changes in legislation have shortened the time when records are kept by the agencies and made it possible to transfer electronic records to the national archives as soon as the agency no longer has a need for them. If the archive is to preserve all the significant properties of a record, the record must reach the archive with all such properties present and their existence recorded in the metadata. Shortening the transfer deadlines will certainly improve the chances that the archive will receive an "original" electronic record and as much metadata with it as was created with the record.

Many archives have developed semi-automated tools for transferring electronic records and their metadata from agencies to the archive. These tools serve several purposes, including the assessment of significant properties of records, ensuring compatibility of metadata and performing quality checks of records. The National Archives in the UK ran the DRUID (*Departmental Record Users Information Database*) project to standardise the description of records as well as applications for and timing of transfers. Using a unified database reduced significantly the formerly paper-based administration of the records transfer process and in the National Archives internally. The DRUID software has by now been upgraded with new tools that connects it directly to the National Archives automated cataloguing system PROCAT to which government departments can enter descriptions of their records directly.[68]

The Norwegian records management standard NOARK includes procedures for transferring records from agencies to the National Archives. The National Archives has also developed a number of automated tools for creating the description of electronic records and the quality control of existing metadata. Perhaps the most important of these tools is the ADDMML (*Archives' Data Description and Manipulation Mark-up Language*),[69] an XML DTD that is used for describing the records to be transferred and that forms the basis for processing of records within the National Archives with other automated tools (e.g., Arkadukt, Arkade, etc.).[70]

A similar model for describing electronic records has been set up by the Danish National Archives where the agencies are filling in an XML-based description form developed by the archives.[71] The description model is also available in MS Word[72] and a detailed user manual has been published to assist the agencies filling in the description forms.[73]

Several countries have set up Government Information Locator Services (GILS). These serve, albeit indirectly, the same purpose – the metadata standards effectively established through GILS mean that records generated in agencies are described with essential metadata. These

---

[67] cf. OCLC/RLG, *Trusted Digital Repositories: Attributes and Responsibilities* (2002), p. 24
[68] cf. http://www.nationalarchives.gov.uk/recordsmanagement/advice/cataloguing.htm
[69] http://www.arkivverket.no/noark-4/dtd/addmml.dtd
[70] cf. J.A. Haugen, *Arkivenes plass i den informationsteknologiske utviklingen* (2000), pp. 133-138
[71] http://www.sa.dk/sa/itogarkiv/teknologi/vejledning/sa_md_11.xsd
[72] http://www.sa.dk/sa/itogarkiv/teknologi/vejledning/XMLMacro.dot
[73] http://www.sa.dk/sa/itogarkiv/teknologi/vejledning/metadatatest.htm

metadata can be used by archives to appraise, transfer and assess the processing needs of records in agencies. Some examples of governmental information locator services and associated metadata standards are:

- Australian AGILS,[74] where one of the main developers is the National Archives of Australia;
- UK *Electronic Government Interoperability Framework* (e-GIF)[75] and the *e-Government Metadata Standard* (e-GMS),[76] where the National Archives is involved in the development of metadata standards;
- United States *Government Information Locator Service* (GILS);[77]
- Italian government agencies unified "net" and records registration standard *Rete Unitaria della Pubblica Amministrazione* (RUPA);[78]
- New Zealand *Government Locator Service* (NZGLS);[79]
- etc.

Due to different legal systems the deadline for transferring records from agencies to the national archives differs from country to country, but the average is 25-30 years after the records were created. This length of time is clearly too long for electronic records – the records will be unusable by the time they reach the national archives if the agency itself does not get involved in pro-active digital preservation. In order to get a better overview of electronic records held by the agencies and to plan better for their processing, national archives are changing legislation to shorten the period of electronic records storage at the agencies. Agencies should be able to transfer their electronic records to the national archives sooner than the 25-30 year deadline. The average period after which agencies can transfer or deposit their records to the national archives is now becoming only 5-7 years after their creation. Changes in legislation to allow for this have been discussed for example in Italy and the UK. Australian National Archives, who for several years was pioneering the non-custodial method of preservation for electronic records, has changed its policy since 2000, has launched two new programmes *e-Permanence* and *Agency to Researcher (AtoR) Digital Preservation Project*,[80] and is now receiving electronic records for long-term preservation.

In Scandinavian countries, depositing of records in the national archives is common practice, where the archives takes the responsibility for keeping and preserving the records, but the agency has the right to borrow or remove its records from the archive. Regulations for this practice have been issued by the national archivist in Norway[81] and Denmark.[82] In Sweden, where agencies and companies can legally deposit their paper records with the National Archives, no special arrangement has been made for the electronic records. [83]

3.4 Conclusion

[74] cf. http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html
[75] http://www.govtalk.gov.uk/schemasstandards/egif.asp
[76] http://www.govtalk.gov.uk/schemasstandards/metadata.asp
[77] cf. http://www.archives.gov/records_management/policy_and_guidance/gils.html
[78] cf. http://www.ctrupa.it/RETE-RUPA/Generalit-/index.htm and http://www.ctrupa.it/allegati/rupa/Schema_Interop_IndicePA-1.pdf
[79] http://www.e-government.govt.nz/nzgls/standard/index.asp
[80] http://www.naa.gov.au/recordkeeping/preservation/digital/summary.html
[81] *Forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver* (2002) http://www.lovdata.no/for/sf/kk/xk-19991201-1566.html
[82] D. Tørning, *Handling of the electronic records issue and cooperation with public administration. Experience of the Danish National Archives* (1997)
[83] *Arkiv för alla – nu och i framtiden* (2002), pp. 100-101

Preserving something digitally has and still does mean that some properties or functionalities of the preserved object are eventually lost either through conversion or changing technology. Attempts to arrive at a theoretical definition what or what kind should an electronic record be, so that it could be preserved forever without losses, have not produced final results yet. Attempts to define archival file formats in practice, or define descriptive "wrappers" around records have also not given widely accepted results as yet. Work on brining closer to each other and connecting the digital object used in daily management of records and objects fit for preservation for long term is continuing.

Although archives in some countries are beginning to benefit from the general e-government developments and programmes that establish standards for metadata, records management and information exchange, the prime responsibility for good, well described records reaching the archives, lies still with the national archives.[84]

As part of their overall duties and services, archives have started to develop tools for automatic or semi-automatic transfer of electronic records and records metadata from agencies, quality checking of transferred records and creation of their description. Since the role of the creating agency in ensuring the quality and authenticity of records has risen significantly with the use of electronic records, it is also in the interests of archives to provide the creating agencies with tools and guidance for creating and maintaining quality digital records. One such measure is the archives taking the responsibility for records preservation at a much earlier date than the traditional transfer deadlines of 20-30 years. Archives are increasingly taking on the role of the mediator "between the agency and researcher".

[84] cf. M. van Dijk, *It Always Hurts the First Time: Experiences with Transferred Electronic Records* (2003)

4. Digital preservation and digital archive

Digital preservation is still regarded as an issue that is not completely resolved, yet it is, nevertheless, being put into practice in archives, its methods are discussed and techniques compared. Perhaps we do not have a complete disquisition of the digital preservation problem yet, at least partly because the theory of digital preservation had to be "invented *post factum*", when archives had already been preserving digital material already for quite some time and the information technology specialists had already developed methods for it. Digital technology is developing at a rapid pace and it is difficult to come up with fundamental archival treatment of all the issues raised by the use of this changing technology. This is probably the main reason why so many research projects are dwelling on smaller, individual problems within the digital preservation domain. This "bottom-up" method of building a theory of digital preservation has been compared to building a house from individual bricks[85] (but it remains unclear whether there is an overall design and drawings for the house that is being built) or to solving a jigsaw puzzle by trying to connect the individual pieces.[86] One way or the other, there seems to be an implicit assumption in the discussions of digital preservation that the "whole picture or drawing" does indeed exist and we just have to find it, invent it, or recognise it.

The theory of digital preservation was, for a number of years "stuck" on the level of physical storage media and bit streams, but has developed further to consider the logical, or the software level issues of preserving access to digital objects. In recent years, the central focus of research projects and theorists has shifted to the 'conceptual level', or the digital object level (e.g. a record) and defining requirements for digital preservation based on the real-life objects that the archivists are handling in their daily job. Significant properties of an electronic record, its use and context all have a bearing on the preservation requirements of the two more basic levels (the logical and physical levels). As a result, the former theories that in order to preserve the electronic record, its constituent file and bit-stream must be preserved intact, are being revised and the complex relationships between the physical, logical and conceptual level of a digital record are being studied further. Any digital preservation strategy must take into account all three levels of a digital object and significant properties of the object on these three levels.

For the past five-six years the main discussions within digital preservation domain have been largely about two methods for digital preservation: emulation and migration, and their respective benefits. The emulation strategy is aiming to preserve the possibility of running the original software for using the preserved electronic record or other digital object as it appeared originally. Migration, on the other hand, will change the logical level encoding of the digital object so that it can be used with the current software and hardware. While emulating the hard- and software environment rests on the premise that it is the original technical environment that helps preserving an authentic record, the migration strategy is based on belief that it is not necessary to preserve the physical and logical level in order to preserve an authentic electronic record, but they can, under certain conditions, be changed. In addition to these two strategies, new approaches have been suggested in recent years and these will be analysed in the following.

[85] G. Mackenzie, *Searching for Solutions: Electronic Records Problems Worldwide* (2000)
[86] M. Hedstrom, *Digital Preservation: Problems and Prospects* (2001)

4.1 Digital preservation strategies

4.1.1 Emulation

It became clear, quite early on in the development of digital archives, that establishing a so called museum of computer technology is not a viable long-term strategy for digital preservation for costs reasons. Emulation is, in a way, an extension of the museum of computer technology idea, aiming to achieve preservation of the functionality of the original technology through artificially simulating it under a newer technological environment.

Theory of emulation

Emulation means the 're-creation on current hardware of the technical environment required to view and use digital objects from earlier times'.[87] Technical environment that may need to be emulated may include not only the original hardware, but also operating systems and application software. Emulation is seen as offering a method of preservation that can preserve the functionality and the 'look and feel' of digital objects that migration may not be able to achieve; and it could prove to be much more cost effective solution in certain circumstances — creating one emulator can be cheaper than migrating every digital object in the archive.

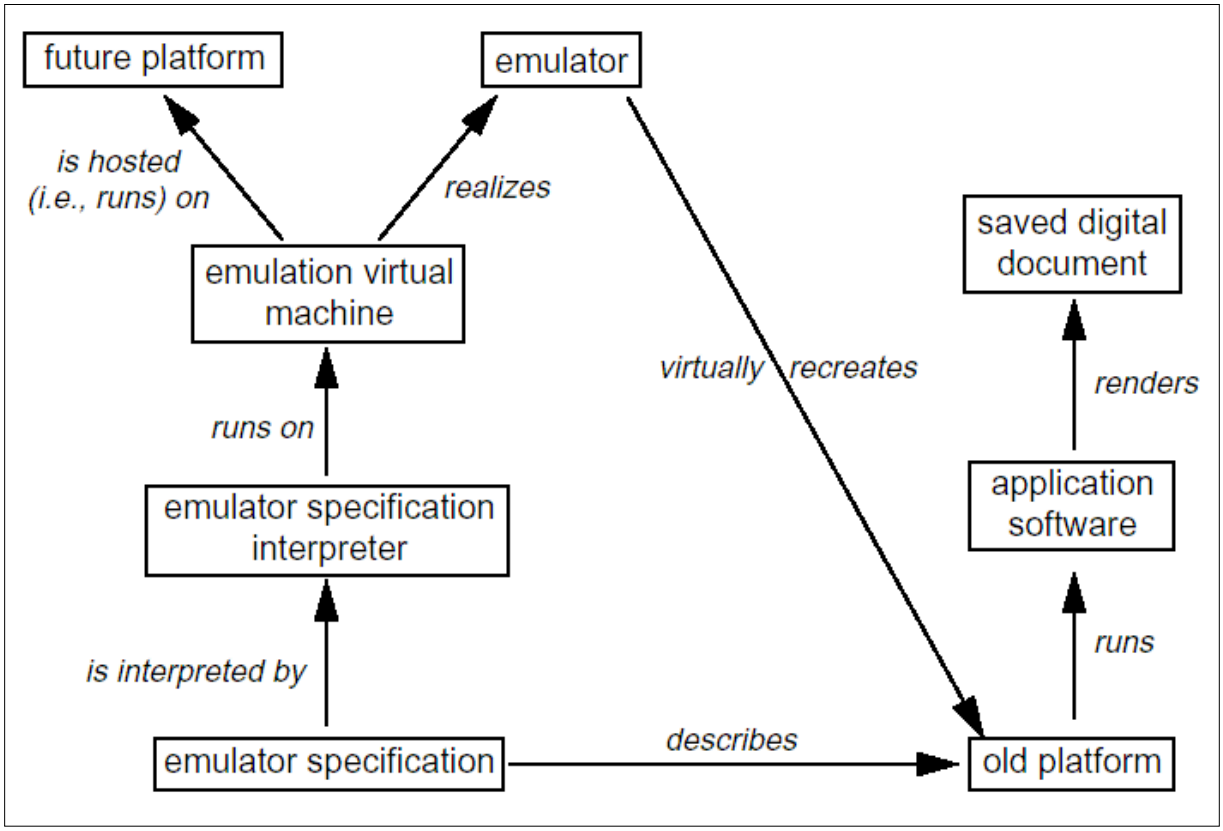

Figure 1. Emulation strategy for digital preservation (according to J. Rothenberg).

Jeff Rothenberg, perhaps the best known advocate of the emulation strategy, has suggested to work on developing emulator specifications that would make it possible to build an actual emulator when the need arose. The "ideal emulation solution" has been defined by him as:

> "[A]n ideal approach should provide a single, extensible, long-term solution that can be designed once and for all and applied uniformly, automatically, and in synchrony to all

[87] D. Holdsworth, P. Wheatley, *Emulation, Preservation and Abstraction* (2001)

types of documents and all media, with minimal human intervention."[88]

This is, in essence, an attempt to specify a single approach to emulation that will work for all possible cases. Rothenberg has developed his idea further,[89] but at the same time this approach has been the most criticised, also by other proponents of emulation. The critics of emulation strategy have pointed out that:[90]

- So far, the strategy has largely existed in theory alone as there has been very little practical experience with emulators;
- The cost-effectiveness of developing an emulator has not been demonstrated conclusively;
- Emulation is trying to preserve the wrong thing by preserving information systems functionality rather than records;
- Emulation suffers from the same complication as migration sometimes does: the lack of reliable software and systems documentation and specifications that are required for building emulators;
- Preserving and using such proprietary specifications can raise intellectual property and copyright questions;
- Components of a stored digital objects may represent software that was developed over a period of time, because software developers re-use old components when releasing new versions. Therefore, an object can be assessed as to the age of its components.

Most implementations of the emulation strategy in practice have ended up defining levels that can be achieved with emulation (e.g., emulating older hardware; emulating commands of one operating system under another; using older application software on new operating system platforms; etc.), and have distanced from the original Rothenberg's idea of defining a single universal emulator.

The CAMiLEON project findings

One of the most influential research projects working on the emulation strategy was the CAMiLEON project (*Creative Archiving at Michigan & Leeds: Emulating the Old on the New*).[91] As the name reveals, CAMiLEON was a collaborative project between libraries of two universities – Michigan and Leeds; In the case of Leeds University it became the continuation of the CEDARDS project. Started in 1999 and initially planned as a three-year project, it actually finished in September 2003 and the last results are only being published now. The three main objectives of the project were:

- To explore the options for long-term retention of the original functionality and 'look and feel' of digital objects.
- To investigate technology emulation as a strategy for long-term preservation and access to digital objects — evaluate publicly available emulators, explore emulation development, conduct test cases, etc.
- To consider where and how emulation fits into a suite of digital preservation strategies.

[88] J. Rothenberg, *Avoiding Technical Quicksand* (1999)
[89] J. Rothenberg, *An Experiment in Using Emulation to Preserve Digital Publications* (2000)
[90] cf. D. Bearman, *Reality and Chimeras in the Preservation of Electronic Records* (1999); J.C. Bennett, *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Materials* (1997); S. Granger, *Emulation as a Digital Preservation Strategy* (2000)
[91] http://www.si.umich.edu/CAMILEON/

The CAMiLEON approach was not to develop a "one size fits all" solution emulator, but to study different emulators that would solve specific problems. The project was looking at material from the 1970s and 1980s (e.g., the BBC Domesday video disc, the George3 operating system on ICL1900 range of computers, etc.) and showing that it is possible to map from an old architecture to one that is radically different.[92] In the end, only one example was chosen for emulation – the BBC Domesday video disc[93] that was practically fallen out of use because the LV-ROM equipment required to read the video discs and the BBC Master computers, not to mention people who could understand and write code in the BCPL programming language, were no longer available. The CAMiLEON project developed, after a longer period of research, a demonstrator system DomesEm for viewing and using the BBC Domesday discs. The work done by the CAMiLEON project on recovering the BBC Domesday multimedia system has been used and developed further by the UK National Archives where the new system is also available for use.[94]

The cost-effectiveness of emulation as a strategy depends partly on the longevity of the developed emulator and how easy it is to update it. The longevity of an emulator is determined, among other things, by the initial platform chosen for it and how much non-standard source code it includes. CAMiLEON studies recommend using C programming language as the best solution for ensuring long life of emulators.[95] It is, nevertheless, likely that every emulator will incorporate some portion of non-standard code (e.g., for rendering of the emulated machine's display) and this must be thoroughly documented.[96] The CAMiLEON emulation strategy is envisaged as a stepwise process and open for the case when an emulator becomes obsolescent. When an obsolete emulator is updated to run on a new platform, the preserved digital object must retain its initial byte-stream throughout the emulation cycles.

The CAMiLEON project publications see emulation as a part of an overall preservation process and not as the *only* viable solution to digital preservation problem. Compared to other preservation methods, emulation has a potential of being able to preserve not only the content, structure and processibility of a digital object, but also its visual aspects, linkages it may include and the interactive features it may have, all in one. Advantages of this approach become important when dealing with complex and multimedia digital objects.

The final reports of the CAMiLEON project are yet to be published, but their topics are listed as digital preservation strategies, costs of digital preservation strategies and collection management. A number of CAMiLEON articles and presentations are already available.[97]

The NEDLIB project

The emulation strategy has been tested by several research projects in the Netherlands. Inspired by J. Rothenberg's ideas, the NEDLIB (*Networked European Deposit Library*) project was to test various emulation techniques. The project that ran from 1998 to 2000, led by the Dutch Royal Library, developed an architectural framework for what it called a deposit system for electronic publications (DSEP) that was broadly based on the OAIS model. The analysis of preservation strategies had the aim of testing and evaluating the feasibility of using emulation as a means of preserving electronic publications. Report commissioned from J.

[92] D. Holdsworth, P. Wheatley, *Emulation, Preservation and Abstraction* (2001)
[93] http://www.si.umich.edu/CAMILEON/domesday/domesday.html
[94] http://www.nationalarchives.gov.uk/preservation/research/domesday.htm/default.htm
[95] D. Holdsworth, *C-ing ahead for digital longevity* (2001)
[96] S. Granger, *Digital Preservation & Emulation: from theory to practice* (2001)
[97] http://www.si.umich.edu/CAMILEON/reports/reports.html

Rothenberg[98] provides an overview of the first iteration of the experiment that was conducted to test the experimental environment and procedures themselves, using existing commercial emulation software. Based on a selection of digital publications, authenticity criteria were developed that need to be preserved for each type of publication (e.g., functionality, behaviour, content). Based on these criteria, aspects of hardware were identified that need to be emulated in the future and this, in turn, allowed to make recommendations for metadata elements that need to be recorded and preserved in order to make the development of emulators possible. Emulation programs of MS Windows operating system were then used on a Macintosh computer to test the set criteria and requirements. The results were evaluated as 'surprisingly encouraging' and criteria for validation tests of the emulation results were developed and documented in the report. Further tests of emulation were planned and negotiations held with the Dutch Royal Library and the National Archives, who had also commissioned a report from RAND Europe on digital preservation strategies,[99] but no results of these tests have been published.

Based on these experiments, the Dutch *Digitale Duurzaamheid* project, where both the Royal Library and the National Libraries participated, began to investigate the *Universal Virtual Computer* (UVC) concept developed by the IBM Research Laboratory.

The Universal Virtual Computer method

The UVC is one variant of the emulation strategy that was first proposed by Raymond Lorie of the IBM Almaden Research Centre that has been running a Long Term Archival research project since 2000. In essence, the UVC takes further J. Rothenberg's attempt to define a 'one size fits all' solution to maintaining continued access to digital objects.[100] The suggested strategy addresses the problem of interpreting data files in the future by writing a program to carry out this interpretation automatically, in the machine language of a Universal Virtual Computer (UVC). Such a program would be written at the time when the record was archived and would be preserved together with the record itself. The program runs on what is called a UVC Interpreter (a virtual machine). In order to interpret the record on a future computer, the UVC Interpreter is required and it should be produced from the specifications of the UVC.

The universal virtual computer should be defined so as to be able to carry out software functions on different platforms (e.g., like the Java language that, however, has not been written for preservation needs). The proposed approach makes a distinction between data archiving, which does not require full emulation, and program archiving, which does. Archiving a document or a data file requires the archiving of description of the use and purpose of the file, as part of the metadata and a UVC program together with it. UVC makes use of a conversion program capable of decoding the original form of the data into a logical format that will be much easier to understand in the future. Unlike in the case of conventional emulation, this conversion program is written today (for a UVC machine), not in the future, when the emulation needs to be used for preserved material. For programs, the UVC-based methodology will rely on a UVC emulator for the present computer technology written now, and a UVC emulator written for the future computer platform. This differs from the emulation method proposed by J. Rothenberg in that it does not require writing an emulator for a real computer from the past in the future. The program to be archived and its data files or documents would essentially be archived with an emulator written in the UVC machine language. Such an emulator would specify the behaviour and the functioning of the original

[98] J. Rothenberg, *An Experiment in Using Emulation to Preserve Digital Publications* (2000)
[99] J. Rothenberg, T. Bikson, *Preservation: Carrying Authentic, Understandable and Usable Digital Records Through Time* (1999)
[100] R. Lorie, *Long Term Archiving of Digital Information* (2000)

program and computer. In the future, the UVC Interpreter interprets the UVC instructions that emulate the old instructions; that emulation essentially produces an equivalent of the old machine, which then executes the original application code. The execution yields the same results as the original program but ideally, the retained metadata should contain a user's guide explaining how to run the program.

What the UVC method accomplishes is to decrease the dependency of applications on proprietary file formats and the need to convert to standardised formats. For any new format appearing, writing the UVC program to decode it will ensure the preservation of its data. The method also favours doing the preparatory work now — when the information is well known — rather than postponing it to the future when the difficulty of describing the functionalities of systems would be much greater. However, the strategy relies on the developer of a new file format producing a program able to decode the data structure and to return the data according to a logical structure. That program can then be compiled onto UVC and packaged for archiving. Such an agreement between all parties involved (developers of new technologies, systems and file formats) has not yet come under wider discussion. So far, the UVC can offer preservation of only a limited number of core functions of a software package – software that can be run on any platform can not be fully compatible with all of these platforms.

The UVC concept has been tested in practice by a collaborative *Long Term Preservation Study* project of the Dutch Royal Library and the IBM. The project included defining functionalities for a digital preservation system (DIAS) to be capable of utilising the UVC method. The first tests were carried out with the PDF files that were described in the UVC language and their use simulated on a different computer platform.[101] The process involved creating raster images of every page of a PDF file and extracting the text on the page to be saved separately. A full-text search engine built for using such an archive allowed finding the search-word in the text and displaying the image of the corresponding page through viewer software. At present, a separate viewer is needed for each type of the UVC encoded data file. In the next phase of the UVC development, classes of objects will be formed that behave according to the same logic. A class of objects like this (e.g., files in different image formats) will produce one UVC encoding, for which only one viewer will have to be developed. It will, however, still be necessary to develop an individual UVC data format-decoding program for each of these file formats.

The Dutch National Archives is carrying out further tests of the UVC method on different types of electronic records (e.g., text, spreadsheets). The first results that are being published are not very encouraging on the grounds of the lack of reliable documentation for the proprietary file formats.[102] The Dutch Royal Library and IBM have recently concluded the *Universal Virtual Computer for JPEG* project that developed an operational 'permanent access tool' that makes the viewing of JPEG image platform independent.[103]

Emulation – conclusions

On the level of national archives the emulation strategy has only been considered as a theoretical option (e.g., Sweden, Great Britain, United States). The strategy has been properly tested in practice only in the Netherlands[104] where the National Archives has so far ingested only a few digital records. It is likely, however, that emulators will find a wider use in the future, especially for the preservation of complex file formats and software systems that

[101] R. Lorie, *The UVC: a Method for Preserving Digital Documents - Proof of Concept* (2002)
[102] J. Slats, R. Verdegem, *Practical experiences of the Dutch Digital Preservation Testbed* (2004)
[103] http://www.kb.nl/kb/resources/frameset_kb.html?/kb/hrd/dd/dd_onderzoek/uvc_voor_images-en.html
[104] cf. Digital Preservation Testbed, *Emulation: Context and Current Status* (2003)

contain multimedia, virtual reality, simulations and also binary executable files. For the time being, the unresolved issues with emulation include what metadata is required for creating an emulator – would it be possible to create a new emulator when the need for it arises in the future based on the standard digital preservation metadata, or should one start amending these metadata already now with relevant information for emulators? As a preservation strategy, emulation is directed at preserving and simulating the functions and properties of the technical environment of a digital record, but what kind and how thorough knowledge of the system to be emulated is required to create an emulator? Answers to these questions are, most likely, dependent on the uses of the emulated record in future but, unfortunately, these are difficult to predict.

After first tests of the strategy the theoreticians of emulation, too, reached the conclusion that building emulators for all types of digital objects is not feasible and cost-efficient approach. Alternatives have been offered through defining emulators for different levels (e.g., hardware platform, operating system, application software) and their use according to the need either individually or together; or building one universal emulator that can be applied in the future to more than one file format or type of object. *Universal Virtual Computer* theory is an extension of this latter approach, although it also includes elements of the migration strategy. First tests show that the UVC method and its variants may be applicable in a digital archive that is preserving a large variety of different file formats and where the usage requirements of preserved objects are limited (e.g., reading, printing). The larger the variety of file formats in the repository and the more future usage situations need to be counted with, the more complex the UVC interpreter will need to grow and together with it the volume of metadata describing each preserved object. However, most records that have been produced to date are not inherently digital and do not necessarily demand the preservation of their appearance and behaviour that are offered by emulation. For example, simple text or page image records may be adequately preserved (at least for most purposes) by less complex approaches that emulation.

Testing the emulation strategy in practice has demonstrated the feasibility of this approach, but the cost of developing emulators has been high albeit, admittedly, within the context of research projects. Although the use of emulation to preserve digital records still requires resolving some significant issues, its potential low cost, universality and ability to preserve originals along with all of their inherently-digital aspects, argue that it is well worth pursuing.[105] It will take probably another 5-10 years before the emulation strategy can safely be applied to preserving electronic records with permanent value; by this time the theory of emulation will have developed further and archives will be acquiring more and more complex digital objects that truly need emulation in the future.

4.1.2 Migration strategies

Theory of data migration

Migration strategy is based on conversion of files and does not aim to preserve or simulate the functionality of electronic record's technological platform. Instead, the migration changes the logical encoding of electronic records (usually their file format) so that they could be used with current hard- and software when the original technical environment has become obsolete.

[105] *ibid.*, p. 56

There are a number of definitions of migration but the most widely cited one is from the 1996 report of the Task Force on the Archiving of Digital Information:[106]

> "Migration is a set of organised tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation."

This broad definition leaves room for a variety of processes carried out as part of migration and indeed, a number of different subtypes of migration have been identified, although in essence they all mean changes in the encoding of information. Charles Dollar has defined the components of a migration strategy as following:[107]

> "**Reformat:** causes a change in the underlying bit stream of electronic records when the physical carrier is changed, or the code is transformed, but without any alteration in the physical representation or intellectual content.
>
> **Copy:** transfer electronic records from old storage media to new storage media with the same format specifications without any loss in structure, content and context.
>
> **Conversion:** automatic transfer of electronic records from one application environment to a new application environment with little or no loss in structure and no loss of content or context even though the underlying bit stream is altered.
>
> **Migration:** conversion that requires specific tools where neither backward compatibility nor export/import gateways exist between the legacy system that contains the records and the new application system."

Although correct from the technological point of view, this narrow definition has not found wider use. In Ch. Dollar's view the need for migration, as defined by him, will decrease in the future because software developers will use more and more standard formats that are more interoperable and simple conversion will suffice for preservation of records.

The ISO 14721:2003 standard "Open archival information system – Reference model" ch. 5.1.3 gives the following types of migration (in the increasing order of danger of information loss through conversion):

> "**Refreshment:** A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by copying the bits on the medium used to hold AIPs and to manage and access the medium. […]
>
> **Replication:** A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication may require changes to the Archival Storage mapping infrastructure.
>
> **Repackaging:** A Digital Migration where there is some change in the bits of the Packaging Information.
>
> **Transformation:** A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content."

---

[106] CPA/RLG, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (1996)

[107] Ch. Dollar, *Authentic Electronic Records: Strategies for Long-Term Access* (1999), pp. 59-72

While the first three in this list are essentially changes of the data carrier (i.e., storage media) and belong to the day-to-day practices of any digital archive, then the last form – transformation – means the actual data migration. The OAIS model also makes a distinction between the *reversible migration* and *non-reversible migration* that are used as indicators of success of the migration process (i.e., will reversing the conversion process give the same original result or not).

Different forms of migration that have been put into practice are the following:

Transforming to the so-called permanent carrier (e.g., paper, microfilm): changing the storage media of a digital object to a static carrier which means significant losses in the original use and functionality of the object. Rarely used in archives.

Conversion using backward-compatible software: frequently used for converting between different versions of the same software product (also known as 'version refreshing'). Can be used for records with short- to medium retention periods and that have been created with widely used software. Although interoperability of software systems and backward compatibility functionalities have improved over the years, it is not prudent to rely a long-term preservation policy on backward compatibility alone, because this functionality is at the mercy of software developers alone.

Converting from one file format to another: is based on the interoperability of software packages, but can cause severe losses in the original functionality of the digital object, especially in the case of complex objects (e.g., CAD/CAM, GIS records, etc.).

Converting to standardised file formats: reducing a great variety of different file formats to a limited number of standardised file formats makes their preservation easier and cheaper. Standardised file formats are more stable than proprietary formats and the methods of encoding of information in them are public (e.g., UNICODE).

The CAMiLEON project has developed another classification of individual migration strategies based on their use. The approach recognises that different digital objects have different preservation requirements, and thus proposes the following levels of migration: [108]

"**Minimum preservation** refers to preserving a copy of the byte-streams that make up the original object (a copy of the original byte-stream should always be retained in addition to any other migrations that modify the original).

Minimum migration that requires very little technical work – the tasks could be performed by hand or automated using a simple software tool. A possible example of a minimum migration could be a word processor file that is stripped of all but the common ASCII characters.

Preservation migration represents the most basic form of access to the intellectual content, with a non-technical way of preserving some of the look and feel of the original digital object. It has three sub-divisions:

- **Basic preservation migration**. This involves recording screen shots of the software in use. This could be a snapshot of the actual screen display on the obsolete platform, or a photo of the VDU while the digital object is viewed.

[108] P. Wheatley, *Migration – a CAMiLEON discussion paper* (2001)

- **Annotated preservation migration**. A more comprehensive record than the basic preservation migration which includes textual descriptions and annotations describing the function and look and feel of the object in question.
- **Complex Preservation migration**. As above, but using simple additional methods to capture descriptive information about the original object. An example of which could be captured video sequences to record key processes in the use of the digital object (e.g., for a word processor document this might involve recording the sequences of running the word processor application, loading the digital document and scrolling through some of its content).

Recreation is the re-coding of a digital object by hand. With a document this could mean re-typing the text in a current application and adding formatting to match the original. At the other end of the scale, a complex software object could be re-coded on a current platform from scratch.

Human conversion migration uses exactly the same processes of reproducing the function of the object, but some element of the original object (usually the data rather than the software) will be incorporated in the final migrated object. An example would be a human conversion migration of the BBC Domesday object, which combined the original image and textual data with a newly recreated or re-coded software front end.

Automatic conversion migration uses a software tool to interpret and modify a digital object into a new form. A typical example would be to take a word processor file from the BBC and output the object as a MS Word 98 document."

As can be seen, despite a fairly long period of implementation and use, there still is no unified agreement or understanding, what migration as a preservation strategy entails and how should its terminology be used. At the same time it indicates that data migration is a developing and large discipline where new definitions and new methods can still be added. It is very likely that in the end, the use of terminology will be determined by the practice of archives once they start treating larger quantities of digital objects. Migration research thus far has mainly focussed on individual file formats, data types or types of records. Analyses of migration as a complete strategy have been undertaken seldom.[109]

Migration research

Using migration to preserve electronic records requires thorough studies of both the original and the target file formats to ensure that all significant properties of the record get carried over during conversion to the next format. Precisely this kind of analyses and studies have dominated the data migration research, offering estimates of risks and benefits of using different file formats.[110]

This research has been very useful in informing the decisions made by national archives as to which file formats to accept for ingest from agencies and the decisions for conversion paths in archives that store digital objects in multiple file formats (e.g., ingest format, preservation

[109] Digital Preservation Testbed, *Migration: Context and Current Status* (2001)

[110] cf. J. Coleman, D. Willis, *SGML as a Framework for Digital Preservation and Access* (1997); S. Gilheany, *Expected Usable Lifetime of Different Electronic Formats* (2000); J. Ockerbloom, *Archiving and Preserving PDF Files* (2000); *XML and Digital Preservation* (2002); S. Thomas, *File Formats for Electronic Text* (2002); etc.

format, dissemination format).[111] These published lists of "safe" file formats are gradually becoming the best practice of data migration and have been copied by other institutions that are only at the beginning of building their digital preservation services. This kind of best practice is, however, not always founded on solid theoretical or even empirical grounds, which means that for safe conversions between file formats the associated risks should be assessed for each individual case anew.

Migration research projects have collected detailed technical information about various file formats and in recent years that information is being systematised into file format registries. The National Archives of the UK has developed the PRONOM system[112] that has become an information portal about data file formats, their supporting software products and software requirements of file formats. PRONOM database holds currently information for over 300 different file formats, but the list is of formats is growing rapidly. Future plans for the registry include providing to users file format specifications, expert opinions about their longevity and recommendations for their migration paths, etc.

A similar initiative has been initiated by the JISC and the e-Science Core Programme in the UK, with the aim of creating an information portal for higher education institutions to support digital preservation with information on file formats and their processing.[113] The first step was establishment of an expert centre – Digital Curation Centre – and publishing a report on publicly available information on file formats.[114]

Another file format registry has been proposed by the Harvard University Library in the US – *Global Registry for Digital Format Representation Information* (or GDFR – *Global Digital Format Registry*).[115] The aim of this project is to collect information about different file formats and present it in the form of the Representation Information of the OAIS model. The project is still looking for funding but meanwhile a small demonstrator system has been built at the University of Pennsylvania Library – Fred: A format registry demonstration.[116]

Migration as a preservation strategy has been criticised for conversion decisions that are made without the knowledge of how these decisions affect the future authenticity of the digital object.[117] It is difficult (some say impossible) to predict what impact each migration stage has on the preservation of significant properties of a digital object – whether the timing, tools and the target format used in a migration stage have been correct or not can only be ascertained in the future when it may be too late. Attempts to measure and quantify risks associated with conversions have produced a risk-matrix of migration[118] and core requirements for software used in migration:[119]

- Read the source file and analyse the differences between it and the target format.
- Identify and report the degree of risk if a mismatch occurs.
- Accurately convert the source file(s) to target specifications.
- Work on single files and large collections.

---

[111] see for example the AHDS services' lists of "acceptable formats" (e.g., http://hds.essex.ac.uk/depguide.asp); cf. N. McGovern, *Preservation storage and processing considerations* (2000)
[112] http://www.nationalarchives.gov.uk/pronom/
[113] http://www.jisc.ac.uk/uploaded_documents/6-03%20Circular.doc
[114] *Survey and assessment of sources of information on file formats and software documentation* (2003)
[115] http://hul.harvard.edu/gdfr/
[116] http://tom.library.upenn.edu/cgi-bin/fred
[117] cf. S. Granger, *Emulation as a Digital Preservation Strategy* (2000); M. Hedstrom, C. Lampe, *Emulation vs Migration: Do Users Care?* (2001)
[118] J. Bennett, *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material* (1997)
[119] G. Lawrence, et al., *Risk Management of Digital Information: A File Format Investigation* (2000)

- Provide a record of its conversions for inclusion in the migration project documentation.

Both of the projects have stressed the importance of file format specifications for assessing migration risks – these should be publicly available or accessible under certain conditions. The cited reports give drastic examples of the speed of file format obsolescence and unreliability of documentation kept by software producers.

Migrating a large digital collection can be time-, labour- and resource intensive, unless some of the work can be automated. Attempts to automate migration have, however, not been very successful. With the information available at present about file formats and software systems, a lot of the research, analysis and testing for automating migration still needs to be carried out manually. If migration is undertaken every time a file format comes under threat of extinction and user requirements for the digital objects are not clearly defined, it is very difficult to calculate the cost of digital preservation based on migration. There are practically no reliable models of cost calculation for digital preservation – the existing models encompass simple and standard file formats, while the everyday reality of archives is much more complex. Some suggestions for calculating the cost of migration-based preservation were made by the CEDARS project[120] and it is to be hoped that the CAMiLEON project will continue with this topic.

Migration on demand

As with emulation, so has the migration theory developed over time and new versions of it are emerging in response to criticisms. One such example is the so-called "migration on demand" concept that has focussed on the conversion tools and software instead of the preserved files. Because the traditional, step-by-step migration carries in it an inherent danger that with every new migration stage an earlier wrong decision or omission is repeated and deepened, it would be relatively safer and cheaper to preserve the original file and migrate it only at the time of use. This, however, requires a constant update, testing and developing of migration tools, to guarantee the lossless compatibility of the original file and the current hard- and software. The CAMiLEON project that developed this concept has tested this approach with some graphics file formats (WMF, Draw, SVG) and demonstrated its success with the help of reversible migration.[121]

Typed Object Model (TOM)

An example of an automated migration tool is the *Typed Object Model* (TOM)[122] system that was initially developed at the Carnegie Mellon University but has since moved to Pennsylvania University Library. The TOM system describes the behaviours and representations of file formats and provides tools for using the file formats. It is based on file format specifications and publicly available conversion tools. The TOM broker (essentially an automatic conversion software tool) treats digital objects based on their 'types' – each digital object is not only a combination of a byte-stream and a file format, but possesses other properties, operations, semantics and classes of methods.[123] For example, an e-mail message is, in addition to its file properties, characterised by conceptual elements like "To:", "From:", "Date:", "Subject:", etc.

[120] S. Granger, et al., *Cost elements of digital preservation* (2000)
[121] P. Mellor, et al., *Migration on Request: A Practical Technique for Preservation* (2002)
[122] http://tom.library.upenn.edu/
[123] J. Wing, J. Ockerbloom, *Respectful Type Converters For Mutable Types* (1999)

In TOM, a format is a type combined with a sequence of encodings that represents the type concretely, and formats can have conversions associated with them. The automated conversion tool contains pre-defined types and classes that can be converted into each-other using specialised converters (e.g., rtf2html, latex2html, etc.) some of which are freeware, some have been developed specifically for the TOM project.
TOM system is open to definition of new types and new type conversion tools, but the technical nature of the systems design makes these options more usable to computer scientists rather than archivists or librarians. File format documentation, tutorials, and software tools are not part of TOM and the system does not function on its own – it need maintenance and updates of type definitions, file format information, etc.
The conversion system is being used by the Pennsylvania University digital library but is also freely available on the web.[124] Although it has been largely a one-man project (designed and developed by John Mark Ockerbloom), the support of the university library and the Mellon Fund will allow to develop the project more systematically.

Standards for interchange formats

Similar to data interchange formats used by different software systems to exchange data, interoperability standards could be used when exchanging data with systems over time. Using standardised file formats for digital preservation resembles the process of transferring data from one system to another, although the future systems have not been defined yet. Methods for capturing and transferring the context and properties of a digital object to a future system have been suggested based on paper and XML.

The *Rosetta Stones Translation* project was initially started in military circles[125] but has recently been taken forward by a research team at the San Diego Supercomputing Center. Like in the TOM project, the idea is to gather information about various data and document types but instead of technical specifications, a representative sample of objects for each type will be collected. The sample must include examples of all significant properties of the type (e.g., a WordPerfect 4.5 document). These properties will be preserved both as a collection of digital files and as a printout on paper or some other human-readable form. In order to use a document type in the future, a converter must be found or built that can convert the document together with all its significant properties to a current software platform. The rules required for building a converter are preserved digitally and on paper or microfilm.
Similar to the migration on demand concept, this approach offers savings in regular migration of whole collections. However, the size of the 'representative sample' certainly needs a clear definition and may vary depending on the type of the object to be preserved. Methods for describing the significant properties and functionalities of the original object are only being developed.

The XML language and standards for marking up contents of documents based on XML are playing an increasingly important role in data interchange between systems. Defining rules for exchange of data between systems is often driven by the requirements of the conceptual level (i.e., which document or information has to be transferred from one system to another). Based on this, the logical objects (including the significant properties) are described and file formats for exchange are determined by these requirements. One such example is the *eXtensible Business Reporting Language* (XBRL)[126] that is used for exchange of financial reports. It is, in essence, an XML language used for creating typical reports on financial status of companies

[124] http://tom.library.upenn.edu/convert/; http://wheel.compose.cs.cmu.edu:8001/cgi-bin/browse
[125] A. Heminger, S. Robertson, *Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents* (1998)
[126] see http://www.xbrl.org

and that can be exchanged between commonly used financial software systems (input/output interfaces have already been defined for these reports in a number of systems). Because the language is ASCII-based and has a so called self-describing XML structure, these documents can be read without special software tools and it is easy in the future to develop software (e.g., XML parsers) for processing the reports. This kind of self-describing documents that include a description of its significant properties are also a type of migration – migration to a standardised format – that will probably be used more often in the future.

Migration – conclusion

Migration is thus far the most common digital preservation strategy. It is best suited for textual documents and other "simpler" types of electronic records, because the changes incurred during conversion are easier to control. Theory of migration was initially borrowed from the information technology practice and has been recently formulated based on archival use of it. Compared to its, perhaps first thorough definition in 1996,[127] the concept is now more precise and includes new techniques to achieve the same end-result. The migration approach has been criticised for its unpredictability and resource demands it sets for preservation – migration presumes a thorough analysis of file formats and records to be migrated before the conversion can be carried out. File formats, their functionality and their significant properties have been the main object of migration strategy research over the past decade. Sufficient information and international experience has been collected by now to begin the definition of automated migration tools and aids, as exemplified by the file format registry projects. Given the slow drive towards fewer and more standardised file formats in the software industry, the migration research will probably diminish somewhat in the future and digital archives will be using and exchanging standardised migration paths stored in registries or automated tools.
Migration has been labelled as an expensive digital preservation strategy, since it requires manual processing of large quantities of individual files. However, based on standardised migration paths it will be relatively easy to develop automated migration tools, one example of which is the TOM system. Since the various file format registries and information portals will be taking on the role of the so called technology watch tasked with assessing the useful lifetimes of different file formats, it will be easier to time the migration paths and conversions in the future.

Migration is and will likely remain the main digital preservation strategy. The research carried out over the past decade has identified the main problems and risk factors associated with data migration and first solutions to minimise the impact of these have been offered. Relating the migration techniques to the requirements of each individual record – its significant properties and the real need for conversion – offers new and cheaper options for using this strategy in practice, yet stressing the importance of risk analysis for digital preservation.

4.1.3 Encapsulation

Encapsulation is understood in the context of digital preservation as preservation of a digital record together with all the information required for successful use of the record in the future. It is a means of overcoming the obsolescence of technological platforms and file formats by storing the information necessary to interpret and use the digital record together with the record in one package. The wrappers or containers used for encapsulation of the record must also include information how the record and the information stored with it (metadata, software specifications, persistent identifiers, etc.) relate to each other.

[127] CPA/RLG, *Preserving Digital Information: Report of the Task Force on Archiving Digital Information* (1996)

The wrapper may include both digital- and analogue material (cf. the Rosetta Stones Translation method described above; sleeve of a storage media that could include important instructions; etc.). Perhaps the most comprehensive list of the necessary information that has to be encapsulated with a record is the definition of the *Archival Information Package* (AIP) of the OAIS model:

- "the representation information used to interpret the bits appropriately for access;
- the provenance to describe the source of the object;
- the context to describe how the object relates to other information outside the container;
- reference to provide one or more identifiers to uniquely identify the object; and
- fixity to provide evidence that the object has not been altered."

The concept of encapsulation is not new (cf. the Bentos container, which was developed by Apple Computers in 1993 to increase compatibility of data between computer applications),[128] but real results for preservation purposes have been achieved in only a few projects.
The British Standards Institute (BSI)[129] developed, in co-operation with the UK National Archives a draft standard IDT/1/4:99/621800DC *Bundles for the Perpetual Preservation of Electronic Documents and Associated Objects* (1999). A bundle is a capsule or wrapper that in addition to the record includes software that is required for using the record, metadata and a description of the whole package.
Encapsulation has been promoted by J. Rothenberg in connection with defining the metadata needed for writing the emulators in the future,[130] but this approach has been criticised by D. Bearman who stated that it is not clear as to how metadata encapsulation strategies may be practically implemented.[131]
The Universal Preservation Format (UPF) method[132] is also based on the idea of encapsulation – the universal format would store in an platform-independent format both the digital object and the metadata that describe it.[133]

Defining the structure and content of the wrapper around a record became significantly easier with the use of XML language. An example of an XML-based wrapper is the *VERS Encapsulated Object* (VEO), defined by the VERS project (see ch. 2.1 above), where the preserved PDF or TIFF file is described by metadata in XML format.[134] No real results of implementations of the OAIS model and its information package concept have been published so far.

The weaknesses of the encapsulation approach are associated with the creation and storage of encapsulated objects – software tools should be able to generate encapsulated records automatically; the potential storage overhead of including documentation about the format within each record wrapper; access to information about unpublished data formats. Even if file format specifications are publicly available, they are often incomplete, including proprietary file format specifications in preservation wrappers may cause problems with the intellectual property rights. No research project has attempted to address all of these issues together, although the same problems are faced by other digital preservation strategies as well.

---

[128] K-H. Lee, et al., *The State of the Art and Practice in Digital Preservation* (2002), p. 98
[129] http://www.bsi-global.com
[130] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* (1999), ch. 9
[131] D. Bearman, *Reality and Chimeras in the Preservation of Electronic Records* (1999)
[132] http://info.wgbh.org/upf/
[133] T. Shepard, *Universal Preservation Format (UPF): Conceptual Framework* (1998)
[134] http://www.prov.vic.gov.au/vers/standard/advice_12/3-1.htm

Encapsulation can also be considered to be a type of migration technique. Although the documentation preserved within the wrapper may delay the need for migration for a long time, the encapsulated information will eventually need to be migrated. Therefore, encapsulation techniques can be applied to the digital resources whose format is well known and that are unlikely to be accessed actively.[135]

4.1.4 Persistent Objects / Persistent Archives

The United States National Partnership for Advanced Computational Infrastructure (NPACI) started some years ago a programme called *Data Intensive Computing Environment* (DICE) with the aim to develop an information architecture to support the creation and to accelerate the publication of scientific data collections. San Diego Supercomputing Center (SDSC) became the main developer of this architecture that requires the integration of distributed persistent digital archives, hierarchical storage systems, databases, data-handling systems, and digital libraries into an integrated scientific information repository. Under the NPACI's DICE programme, and earlier, the SDSC has run several research projects, three of which are directly linked to the persistent archives solution developed for the US National Archives (NARA): the initial co-operation under the DOCT project;[136] the so called NARA projects; and a NHPRC funded project. The NARA and SDSC collaboration has produced a number of concepts and tools for long-term digital preservation that can be summarised as "persistent solutions".

In the first year of research conducted for NARA, SDSC articulated the 'persistent object' approach on the basis of the OAIS model. The persistent object method presumes that all records can be represented as objects with their specific characteristics and behaviour. In the second year of its work, SDSC enriched the architecture and called it 'knowledge-based persistent object preservation'.[137] In parallel a theory of collection-based archives was developed and merging these together it became the 'knowledge based persistent archives' concept.[138]

Persistent objects

In simplified terms, the persistent object is what is created at the ingest stage from a document deposited in a digital archive for preservation – essentially an AIP. For creating a persistent object, all significant properties of the object that is to be preserved are identified and expressed in explicit, abstract (formal) models in XML. Such significant properties would, as a minimum, include a) the internal components of the object; b) the sequence of components within the object; c) the attributes of presentation of the preserved object. Once the properties for a range of typical records have been defined as a model (e.g., the typical structure of an e-mail, a letter, a bill, etc.), these records can be encapsulated in metadata as defined in their respective models. After that, other technical characteristics of records that are dependent on specific hardware or software, proprietary, or subject to obsolescence are to be eliminated by converting to a platform-independent format. Special software 'mediators' are later used to enable future technologies to interpret both the models and the stored metadata for information discovery and delivery, as well as for rebuilding the record collections.[139]

---

[135] K-H. Lee, et al., *The State of the Art and Practice in Digital Preservation* (2002), p. 98
[136] see http://www.sdsc.edu/DOCT/
[137] B. Ludäscher, R. Marciano, R. Moore, *Towards Self-Validating Knowledge-Based Archives* (2001)
[138] R. Moore, *Knowledge-based Persistent Archives* (2001)
[139] cf. K. Thibodeau, R. Marciano, *Building the Archives of the Future: NARA's Electronic Records Archives Program* (2000)

Persistent collection

A persistent object is by default part of a collection – the provenance and original order of records in an archival fond must be maintained by grouping them into collections. Similar to persistent objects, collections are defined at the records ingest stage and their attributes are described in the metadata (e.g., access conditions, curation, index, domain metadata, etc.). The collection has an implied organisation, which is typically a subset of the attributes associated with the digital objects. Links among records and collections are expressed as persistent data values – collections define the context associated with objects that comprise it.
A persistent collection requires the ability to dynamically recreate the collection on a new technology platform (either for access or migration) while maintaining collection's integrity. This is achieved by augmenting the collection with special metadata attributes needed to recreate the data collection at a later stage. SDSC has thus defined collection as an XML view on the original tagged data using an information or context model.[140]

Persistent archive

Persistent archives comprise of both the original digital objects (i.e., records) and the information required to assemble the digital objects into a data collection are archived simultaneously. Digital objects are not archived as stand-alone entities but as members of a digital data collection that can be transferred from one technology platform to another. As the concept developed in the SDSC publications, it became to represent a type of migration mechanism – an interoperability system, able to re-create the archive on a new platform. Persistence is at that demonstrated by dynamically building the data collection from the individual data objects stored in the archive, dynamically creating the relational joins needed to discover information within the data collection, and dynamically constructing the presentation interface for the digital objects.[141] Persistent archive provides the mechanism to ensure that the hardware and software components can be upgraded over time, while maintaining the authenticity of the collection – while the migration occurs, a persistent archive must be able to inter-operate with both the old technology and the new technology.

Persistent archive preservation

Preserving a persistent archive is based on migration strategy, but unlike the traditional method of migration, the basis for decisions to convert the logical encoding of the record is here not the comparison of the original and target formats, but the significant properties of the record itself. The persistent archives method relies on two assumptions: the significant (essential) properties of a record can be specified, and in many cases, a record may have several different, but equally valid materialisations (e.g., representations in different file formats). Generically, NARA defines the essential properties of a record as including its content, structure, context, and presentation. Identifying significant properties does not necessarily entail that these properties cannot change, but that the variability of any property must be specified and not change over time. For example, one of the advantages of computer assisted design is that 3-dimensional digital designs can be rotated and viewed from different angles. The ability of a user to rotate the design defines a variability that must be preserved. If the design were saved in a format which only supported a static display, the result would be considered a version or extract of the original design, not an authentic copy.[142]

[140] cf. R. Moore, et al., *Collection-Based Persistent Digital Archives*, Part I (2000)
[141] cf. A. Rajasekar, et al., *Collection-Based Persistent Archives* (1999); R. Moore, et al., *Collection-Based Persistent Digital Archives*, Part I (2000)
[142] NARA/ERA, *Introduction to Preservation and Access Levels Concepts* (PAL v1.0) (2003)

Records' and records collections' significant properties are described using XML – in the system currently being tested at the SDSC records, their parts and their collections have been tagged to enable subsequent decision-making for preservation. Tags essentially indicate and delimit the atomic "units" that need to be processed for preservation, whereas the content and size (the granularity) of the tags can be variable depending on the requirements set by significant properties. Every tag is linked to one or more higher-level constructs, such as data models, data element definitions, document type definitions, and style sheets defined at the information level, and ontologies, taxonomies, thesauri, topic maps, rules, and textbooks at the knowledge level.[143]

Abstraction and description of significant properties of records and their collections is achieved through data grids. The grid technology maps every new software and functionality of record types and links these to the functions of the digital repository. Current tests with persistent archives rely on a specific systems architecture at the SDSC – the Storage Resource Broker (SRB) – that is compliant with the OAIS reference model. The preservation environment is based on the Storage Resource Broker data grid and links three archives at NARA, the University of Maryland, and SDSC. An important part in the grid technology as well as in the management of a digital repository is played by the persistent identifiers that identify each individual record, software for its use and its grid mapping.[144] Persistent archives based on the grid technology are being actively developed and there are international initiatives for it.[145]

Knowledge-based archives

The knowledge based archive adds the conceptual level information and rules for building models to the collection-based archive. Recording explicitly not just the informational level attributes of the records and the structure of a collection, but the implied relationships that exist between the attributes, the processing steps used to accession the collection and, for example, the preferred organisation of the attributes in user interfaces for viewing the data collection, adds new functionalities to the collection. The knowledge implicitly present in archived objects and collections can entail a number of relationship types (e.g., procedural/ temporal, structural/spatial, logical/semantic, functional, etc.) and can, therefore, be used to facilitate the automated validation and quality assurance of records and archives at the accessioning stage, as well as the infrastructure independent recording of transformations performed at the ingest stage that make it possible to automate the archival reinstantiation process (i.e., access) to a large extent. [146]

In practice, a knowledge-based archive adds knowledge packages (KP) to archived objects (AIP) that capture context known at the time of archiving using logic rules and conceptual models of collections, integrity constraints, derivation rules, virtual relations, etc. Such XML-coded or knowledge representation languages based KPs can be applied at the ingest, migration, and instantiation (access) time. The SDSC has suggested using the ISO 13250 "Topic Maps" standard to maintain mappings between domain concepts and the attribute names of objects. [147]

---

[143] cf. K. Thibodeau, *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years* (2002)

[144] R. Moore, *The San Diego Project: Persistent Objects* (2002)

[145] Persistent Archive Research Group of the Global Grid Forum: http://www.gridforum.org/6_DATA/persist.htm

[146] B. Ludäscher, et al., *Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives* (2001), pp. 6-7; B. Ludäscher, et al., *Towards Self-Validating Knowledge-Based Archives* (2001), pp. 5-8

[147] R. Moore, *Knowledge-based Persistent Archives* (2001)

Persistent archive – conclusions

Although somewhat similar to the migration on demand concept, the method described in this subchapter is different from other approaches to digital preservation. Instead of technological requirements, the persistent object preservation is based on requirements of the record as a conceptual object that includes in addition to technical requirements also other parameters, like authenticity, reliability, context, archival bond, etc. Estimating the risk of platform migration is in this case not just an assessment of file formats and conversion tools, but an analysis of the whole archival collection and its contents. The preserved objects are made platform-independent by a high level abstract description of their properties and functions that are encapsulated in the archive together with its constituent records.

The persistent objects theory is still being developed further, but first results are there. The first co-operation between NARA and SDSC has created a whole research topic that now "lives its own life" and NARA is but one potential customer to the developed solutions and technology. The theory of persistent objects is presented in publications in an abstract, more information technology than archival language, which makes it somewhat difficult to understand to a less technology-competent archivist. Behind the trendy and constantly evolving terminology are, however, relatively simple archival concepts that have been linked to new developments and possibilities of the IT fields.
The archival concepts that are used have given reasons for some criticisms, for example, the concept a digital record is defined as a generic object that can be divided into content and presentation, several undefined presumptions and a somewhat mechanic definition of authenticity (e.g., To prove authenticity, content, context, structure and presentation of a record need to be maintained.). It is still not clear whether XML will survive the tests all the different applications it is being used for, also when describing properties and functionalities of digital records. Although persistent archives are not based or dependent on XML but currently use XML as the best available technology and there is nothing in the architecture that would preclude using other implementation methods in the future should they prove superior.
Some of the criticisms have been taken into account as the theory develops. For example, the relatively mechanical definitions used in the early days (context = provenance, original order = archive, etc.) have been developed further in the knowledge based archives concept. The result is a somewhat complicated version of the archival theory in IT language that is being and will be polished for some time yet. The general approach to the digital preservation problem, is however, fresh and promising.

4.2 Comparing digital preservation strategies – a small summary

Digital preservation and its strategies have developed enormously over the past twenty years – from intuitive practical solutions to a solid theoretical subject or discipline that has its theory and several competing methodologies. Most notably, the development has led away from the preservation of technology and focussing on preservation of records. Publications of the past few years are almost unanimously calling for making preservation choices based on the properties of the object that is to be preserved and not on the requirement of preserving adequately the functionality of the original technology that in real life has very little archival use. There is, as yet, no single long-term digital preservation strategy that would suit for preserving all archival objects – the choice of methods for preserving each individual type of object and record must be based on the assessment of its properties and their requirements. Determining the significant properties of a record is, therefore, necessary not only to establish the record's authenticity criteria and to define the ways of using the record in the future, but also to choose the appropriate preservation method for the record and how this method should

be applied. The mainstream of digital preservation research is currently focussing on generalising the last two of these issues and ways of automating them, because it would be too costly to manually analyse each archival record and its significant properties manually. K. Thibodeau has presented the current digital preservation methods on the following graph:

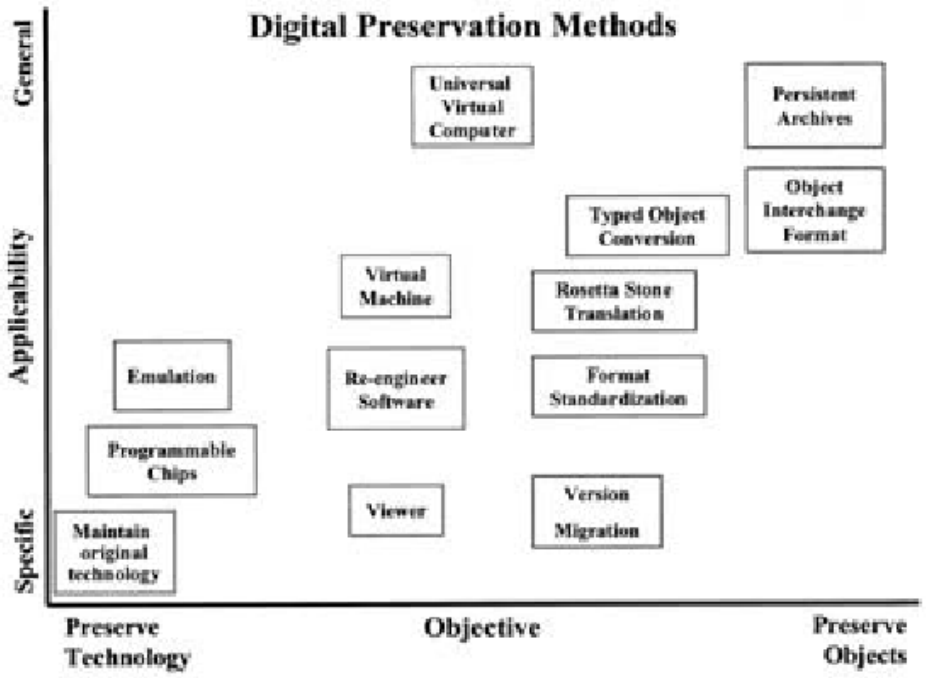

Figure 2. Comparison of digital preservation strategies (after K. Thibodeau).

The long principal discussion between the proponents of emulation and migration strategies has by now abated to a constructive dialogue where neither camp seems to consider their approach the only possible solution. Aside these two strategies, new methods have been proposed and also put into practice – like encapsulation (cf. VERS VEO), TOM, persistent archives. Migration strategy that has been defended as the technology-independent solution has now moved more towards the middle on the scale of technology dependence (see Figure 2) whereas new methods offer more formalised decision making for preservation where technology is only one factor to be considered.

Migration, or more exactly, one of its variants – conversion to standardised file formats – has been the most widespread preservation method so far. It has become clear that this strategy does not offer a complete solution to preserving the usability of all types of digital records over the long term and other strategies will have to be used, too. For archives this constitutes above all a practical problem – how and with the help of what should one manage the growing volume of digital archives? Can the current digital archive systems and tools handle a larger variety of file formats and preservation techniques? What about metadata necessary for digital preservation? The evolution of digital archive models and systems is the topic of the next sub-chapter.

4.3 Digital archive models and systems

For the first thirty years of digital preservation, archives managed their digital collections with the help of a repository for storage of off-line media, or a simple storage system and a catalogue box or a catalogue database. Although the fundamental design of a digital archive system has remained the same — data storage plus metadata database — a contemporary digital archive needs more than a storage area for magnetic tapes and a spreadsheet for the catalogue. The rapid growth of digital material in both volume and complexity, as well as the rising expectations of archives' users for the services they can get and the new emerging digital preservation strategies, have all contributed to the re-definition of digital archive functions. The functionalities and procedures of a digital archive have now been collected into a reference model that has become an ISO standard (ISO 14721). In concert with the development of the (more theoretical) archive models, several organisations have already started implementing new solutions in practice, what could be called digital archive information systems.

4.3.1 The OAIS model

Development of the standard

The Consultative Committee for Space Data Systems (CCSDS) was established in 1982 to provide an international forum for space agencies interested in the collaborative development of standards for data handling in support of space research. In 1990, CCSDS entered into a co-operative agreement with Subcommittee 13 (Space data and information transfer systems) of the Technical Committee 20 (Aircraft and space vehicles) of the ISO. At the request of the ISO, in 1995, CCSDS assumed the task of co-ordinating the development of archive standards for the long-term storage of archival data. Although the CCSDS was initially to address the problems of archiving data obtained from observations of the terrestrial and space environments and used in conjunction with space missions, it soon took an intentionally interdisciplinary view and ensured broad participation in the discussion of a reference model for the long term storage requirements of this digital information.[148] The very first draft of the digital archive model was released already after a year of work;[149] the draft was then discussed in international and national working groups and workshops,[150] which resulted in the publication of the first version of OAIS model in 1999 and its update in 2001. Work had also been started on additional standard guideline detailing the acquisition process: *Producer-Archive Interface Methodology Abstract Standard.*[151]

Development of the reference model was started with the premise that one of the greatest challenges in accepting preservation responsibility within an organisation is finding a shared vocabulary for stakeholders with a variety of backgrounds to use for productive discussion of the issues. Thus, the model was first developed to establish common terms and concepts, provide a framework for elucidating the significant entities and relationships among entities in an archive environment, and serve as the foundation for the development of standards supporting the archive environment. A broader task for the OAIS development has been defined as articulating the functionality and components of any system responsible for preserving any type of information over any length of time. The terminology used to describe the OAIS is often not the traditional archival or recordkeeping terminology since it is intended as a common language within which a diversity of communities can continue to implement

[148] B. Lavoie, *Meeting the challenges of digital preservation: the OAIS reference model* (2000) p. 26
[149] http://ssdoo.gsfc.nasa.gov/nost/isoas/us01/p004.html
[150] cf. http://ssdoo.gsfc.nasa.gov/nost/isoas/dads and http://ssdoo.gsfc.nasa.gov/nost/isoas/awiics
[151] http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf

and develop the OAIS model. The model has certainly been successful in one of its goals — to spur further interest and discussion of digital preservation and archiving issues and standards. The 2002 CCSDS version of the OAIS reference model[152] was proposed and was accepted as an international standard in 2003: ISO 14721:2003 "Space data and information transfer systems – Open archival information system – Reference model".

"Open" archive model

An "open archival information system" is defined in the standard as:[153]

> "an archive, consisting of an organisation of people and systems, that has accepted the responsibility to preserve information and make it available for a designated community".

The word "open" in OAIS refers to the model and future recommendations associated with the model being developed in open forums; it does not make any presumptions about the level of accessibility of information in the archive. The CCSDS began its work by defining what is meant by "archiving of data" and then breaking the "archiving" into functional areas of ingest, storage, data management, administration, preservation planning and access. Further elaboration of interfaces between these functional entities through the informational model and use of "information packages" sets the environment and procedural framework for an archive.

The OAIS environment is structured around four types of actors and the information that they create or require:

- a Producer (a person or system) who produces information to be preserved;
- a Manager who determines the policy framework within which the OAIS will operate;
- the Archive or OAIS itself;
- a Consumer (a person or system) who queries the OAIS in order to find and retrieve relevant information.

A special subset of consumers is the Designated Community who are expected to be able to understand the preserved information. The Management entity does not include the day-to-day administration of the archive, this task is performed by a functional entity within the archive.

The functional model of the OAIS identifies six high-level functions:

- ingest function is responsible for receiving information from producers and preparing it for storage and management within the archive;
- archival storage function handles storage, maintenance and retrieval of archived information;
- data management function co-ordinates the descriptive information pertaining to the archived data and the system information used to support the archive's operation;
- administration function manages day-to-day operation of the archive and develops policies and standards related to the data in the system;
- preservation planning monitors the environment of the OAIS and provides recommendations to ensure that information stored in the OAIS remains accessible;
- access function helps consumers to identify and obtain descriptions of relevant information in the archive and delivers information from the archive to consumers.

[152] http://www.ccsds.org/documents/650x0b1.pdf
[153] *CCSDS 650.0-B-1 Reference Model for an Open Archival Information System (OAIS)* (2002), p. 1-1

The functional entities of the OAIS manage the flow of information from information producers to the archive and from the archive to consumers. Taken together, they identify the key processes in most systems dedicated to preserving digital information, or indeed, to most traditional archives.

The OAIS information model defines the broad types of information that would be required in order to preserve and access an information object stored in a repository. Information itself is defined as any type of knowledge that can be exchanged, that is independent of the forms used to represent it (i.e., physical or digital), and this information is always expressed by some type of data. In order to successfully preserve such a generic Information Object, it is critical for an OAIS to clearly identify and understand the Data Object and its associated Representation Information, which together form what is called an Information Package.

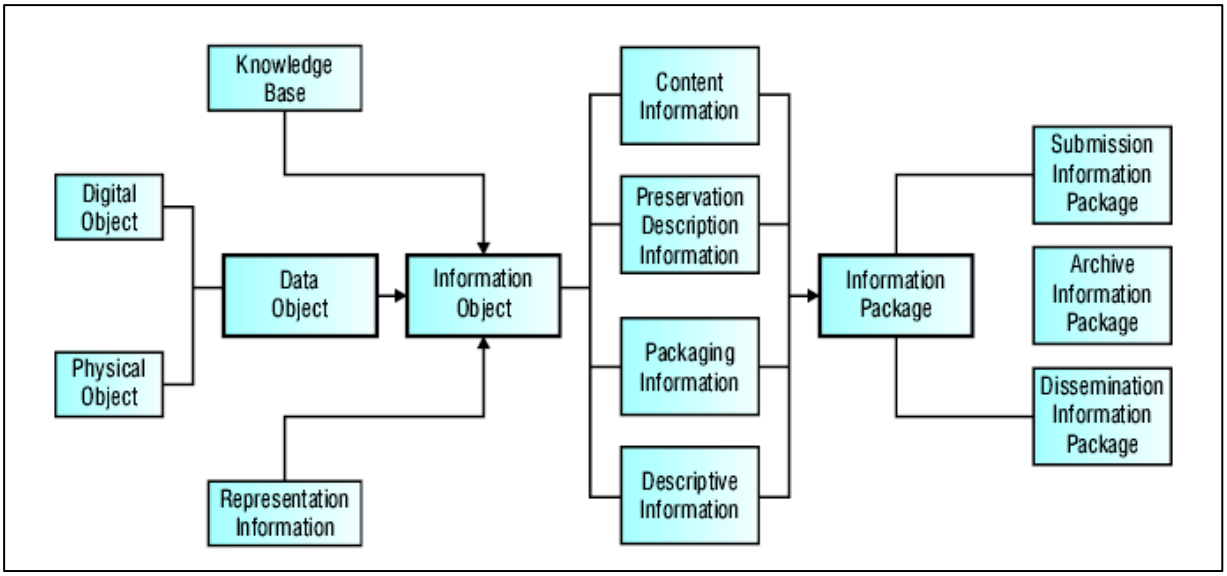

Figure 1. The OAIS information model (after B. Lavoie).

The six functional entities of an OAIS interact with each other by packaging the information they exchange and preserve into specific information packages:

- Submission Information Package (SIP) delivered to the archive by the producer;
- Archive Information Package (AIP) generated from SIP's and stored by the archive;
- Dissemination Information Package (DIP) which is transferred from the archive in response to a request by a consumer.

The information packages contain both the archived file as well as its descriptive metadata. The description is divided into four categories (cf. Figure 1 above): content information, preservation description information, packaging information, descriptive information.

The notion of a generic information package that must be managed through time is fundamental to the OAIS model. The information packages are the means of interaction for players within the OAIS environment and the nature of information packages defines much of the specifics of the interfaces between the functions of an OAIS. For example, the AIP is expected to have all the qualities needed for permanent or long-term preservation of its information object.

The OAIS model is still only an example or blueprint (even in the title the standard is referred to as a "reference model") that does not provide a detailed specification for an archive

implementation. All of the different communities interested in digital preservation will have to apply the model in their own particular contexts, both organisational and technical. The standard does, however, provide a model for assessing and comparing existing archives and their services and functions.

4.3.2 OAIS model implementations and related projects

Libraries were, perhaps, participating in the development of the OAIS reference model more actively than archives, because the need to solve digital archiving issues was more acute at the time in libraries. Several digital archive system development projects were started at the time when OAIS model was being developed, but comparing and commenting on the results of both of these developments came only when the OAIS model was almost at its last version.

The NEDLIB project

The Networked European Deposit Library (NEDLIB) project was initiated by the Standing Committee of the Conference of European National Libraries (CENL) in 1998 with funding from the European Commission as part of its Telematics Application Programme. The project ran until the end of 2000 and it was based on a consortium of national libraries, publishers, information technology organisations and the Dutch National Archive, all led by the Koninklijke Bibliotheek. The overall aim of the NEDLIB project was to construct the basic infrastructure upon which a networked European deposit library can be built and the work was conducted mainly in three areas: architectural framework for a deposit system for electronic publications (DSEP), issues of long-term digital preservation and strategies that would suit for the preservation of electronic publications, building a demonstrator preservation system.[154]

The deposit system for electronic publications (DSEP) developed by the NEDLIB project is broadly based on the OAIS model and augments it by defining a specialised functional entity for preservation.[155] The process of ingest, preservation and use of digital publications was defined as interfaces to a Digital Library System (DLS) — a digital repository that, through specialised interfaces is integrated with traditional library functions. Emulation, which has received very little attention as a preservation strategy in the OAIS model, has been considered as an option in the DSEP system.[156] A guide to technical standards and solutions for implementing a digital deposit system was published in the NEDLIB report series.[157]

Koninklijke Bibliotheek DNEP project

In order to implement the NEDLIB project result, the Dutch National Library (Koninklijke Bibliotheek) began developing its own depository for electronic publications: *Depot van Nederlandse Elektronische Publicaties* (DNEP). After a lengthy tendering process, IBM was chosen to be the partner for developing and delivering the technical infrastructure for a large-scale digital library system. The partnership, known as the Long Term Preservation Study (LTP) is comprised of five sub-projects,[158] that include implementing a repository system that meets the DSEP and DNEP specifications, and conducting further research into digital preservation strategies.

---

[154] T. van der Werf, *Long-term Preservation of Electronic Publications. The NEDLIB project* (1999)
[155] cf. *NEDLIB Contribution to the review of the OAIS Reference Model* (2000)
[156] T. van der Werf, *The Deposit System for Electronic Publications: a Process Model* (2000)
[157] B. Feenstra, *Standards for DSEP: Standards for the Implementation of a Deposit System for Electronic Publications (DSEP)* (2000)
[158] http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study-en.html

The Digital Information Archiving System (DIAS) implemented by IBM at the Koninklijke Bibliotheek is comprised of six modules defined by the OAIS and the NEDLIB project, as well as additional tools and interfaces for adding material to the archive and using it. DIAS is based on IBM DB2 and IBM Content Manager software that run on UNIX operation system, but the user interfaces have been created for MS Windows. Koninklijke Bibliotheek has its own development team working on additional tools for workflow management for the digital archive.[159]

Digital Library Federation

A collaboration project of library organisations and publishers in the United States was initiated in 1999 with the aim of developing infrastructure for preserving electronic scholarly journals. The Digital Library Federation (DLF) consortium[160] is acting as the co-ordinator of altogether seven individual projects between university libraries and publishers.[161] Although other digital preservation models are being studied and tested (e.g., LOCKSS), most electronic journal archives are based on OAIS. Since electronic dissemination of scholarly journals is limited by intellectual property and copyrights, the Digital Repository Service (DRS) systems in development are predominantly focused on preservation, but they must have capabilities of interacting with other library systems.[162]

CEDARS Demonstrator Archive

The Cedars project Demonstrator Archive was set up to test the distributed archive model.[163] It included a number of test sites (university and other libraries and a computing centre) designed to test different aspects of digital archiving with various types of digital materials (image or text files, electronic journals, large online databases and complex multimedia materials) and to illustrate a scaleable distributed digital archive. Cedars distributed archive architecture was based on an implementation of the OAIS reference model. The architecture implemented in the demonstrator functions as a co-operative interworking of three distributed archive stores that share a common namespace with minimally restrictive naming conventions and with indirections allowing the relocation of stored objects between stores without losing addressability via the original name.[164] The aim of the prototype system is threefold: retain the resource in long-term preservation storage; ensure that the preserved digital object can be found; facilitate understanding of the resource.

In the distributed architecture, it is vital that each preserved object has a unique reference – the CEDARS reference ID (CRID) is assigned to object at the ingest stage. Within the facilities for search and retrieval all objects in the archive stores are referenced by their CRIDs. The CRIDs conform to a Uniform Resource Name (URN) syntax and the use of the CRID namespace enables a level of indirection which is crucial to the maintenance of the integrity of representation networks over indefinitely long time. In the Cedars Distributed Archiving Prototype responsibility for the allocation of each sub-domain of CRIDs is held by the local Nameserver custodian. This allows a local archival site to operate a naming policy which suits their local infra-structure and their local collection policies. Each Nameserver custodian is

[159] T. van der Werf, *Experience of the National Library of the Netherlands* (2002)
[160] http://www.diglib.org/dlfhomepage.htm
[161] http://www.diglib.org/preserve/ejp.htm
[162] http://www.diglib.org/preserve/criteria.htm
[163] http://www.leeds.ac.uk/cedars/archive/archive.html
[164] K. Russell, D. Sergeant, *The Cedars Project: Implementing a Model for Distributed Archives* (1999); D. Holdsworth, D. Sergeant, *A Blueprint for Representation Information in the OAIS Model* (2000); D. Holdsworth, *Architecture of CEDARS demonstrator* (2001); *Cedars Guide to the Distributed Digital Archiving Prototype* (2002)

allowed to allocate any name within their domain although they must adhere to the current local naming policy.

In the Cedars Distributed Archiving Prototype each Archival Information Package (AIP) is stored as a single byte-stream containing all of the Preservation Description Information (PDI) and Representation Information and all of the components of the Digital Resource (the Content Digital Object). The format of the internal components of the AIP are as follows:

- Preservation Description Information an XML record;
- Existing Catalogue Records several named byte-streams;
- Representation Information a list of properties (in Java notation);
- Content Digital Object an unmodified byte-stream.

Packaging Information, PDI, and RI are human readable, which provides a failsafe in case the tools to interpret the AIPs cannot be used. The Packaging Information contains a packaging version number and a CRID, and the object referenced by the CRID contains the specification for the structure of the AIP. Representation information contains CRIDs which relate the particular AIP to the objects which form the appropriate Representation Network.
The end-user interface in the Demonstrator Archive is primarily aimed at library staff as users.

VERS digital archive

Victorian Electronic Records Strategy (VERS) for managing electronic records has been discussed above (see ch. 2.1). VERS digital archive is currently in the solution development phase, with functional requirements and technical infrastructure specification for it are available.[165] VERS Centre of Excellence produced a roadmap for replacing the old ArchivesOne system at the Public Record Office of Victoria (PROV) that did not include preservation functions, with a new digital archive system. Specifications produced for the public tender in 2003 outline the connections between the VERS and OAIS models, but the description of VERS digital archive uses archival vocabulary, instead of OAIS terms.

The VERS digital archive will manage and preserve permanent VEOs of the Victorian Government, their quality validation, storage in a digital repository and will enable on-line access to VEOs. The access component to the PROV digital archive will also provide user access to the paper collection at PROV. For paper records, “access” refers to the ability to search or browse the descriptive metadata about records, and to order records for viewing in a reading room. For electronic records, “access” also covers the ability to search or browse descriptive metadata and other metadata about VEOs, as well as rendering copies of VEOs to the user’s desktop.[166] The main task of the preservation function is to manage and preserve the PDF, TIFF and ASCII files that have been encapsulated into XML wrappers and “sealed” with a digital signature. The digital archive is scheduled to be operational for record transfers and for online access in July 2005.

Attributes of a Trusted Digital Repository

A joint task force of the Research Libraries Group (RLG) and the Commission on Preservation and Access (CPA) was defining the nature and properties of sustainable digital archives. The initial project that ran from 1994 to 1996 was later continued as part of the OAIS model development. RLG and OCLC have continued research into digital archives and repositories

[165] http://www.prov.vic.gov.au/vers/projects/Part%20B%20-%20Specification.zip
[166] http://www.prov.vic.gov.au/vers/projects/digitalarchive.htm

also after OAIS has become an international standard. The research is, in particular, seen as important for heritage and scientific data organisations:[167]

> "A definition and consensus is still needed on the characteristics of a sustainable digital archives for large-scale, heterogeneous collections held by research repositories."

The objective was defined as "a rational set of criteria for archives that can hold the full range of digital collections and datasets (including both "born digital" and "born-again digital" information) requiring long-term storage and access systems. The result of the work of an international working group was published in 2001 in the draft report *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*. The final version appeared after a discussion period in 2002.[168]

To ascertain criteria for a trusted digital repository, the working group first studied how trust is established in repositories on three levels:

> How cultural institutions earn the trust of their designated communities.
> How cultural institutions trust third-party service providers.
> How users trust the documents provided to them by a repository

The first of the attributes of a trusted digital repository formulated on the basis of this research, is the compliance with the OAIS reference model requirements. Other criteria include:

- administrative responsibility;
- organisational viability;
- financial sustainability;
- technological and procedural suitability;
- system security;
- procedural accountability.

Thus, a digital repository can earn the trust by demonstrating documented and accountable procedures and services, no requirements have so far been set to the technological architecture.

As a continuation to the definition of attributes for trust, RLG has set up a joint task force with NARA to develop a process of digital repository certification. This process should address the range of functions associated with repositories while providing layers of trust for all involved parties. It should yield a high degree of confidence that the information a repository disseminates is the same information that was ingested and preserved. The certification process must also address the consequences of failure, including fail-safe mechanisms that would enable a certified archival repository to perform rescue of endangered digital information.[169] The task force is currently working on identifying and describing the elements of a digital repository that can be assessed and certified, the first results should be published as a self-certification guide during 2005.

The InterPARES Preservation Task Force digital preservation model

One of the research topics of the InterPARES 1 project (1999-2001) was *Preserving Authentic Electronic Records*, for which an InterPARES Preservation Task Force was formed. The Task Force based its work on the OAIS reference model, with the caveat that the OAIS model is not

[167] http://www.rlg.org/longterm/attribswg.html
[168] *Trusted Digital Repositories: Attributes and Responsibilities* (2002)
[169] http://www.rlg.org/en/page.php?Page_ID=7783

specifically designed to create an environment for preserving only authentic electronic records. Since the requirements for preserving authentic electronic records had not yet been sufficiently clearly defined, InterPARES formed another Task Force to identify conceptual requirements for assessing and maintaining the authenticity of electronic records.
The Preservation Task Force began to formulate preservation business processes from the viewpoint of the person responsible for carrying out actions needed for preserving electronic records. The aim was specifically not to define a system specification or a workflow model to be implemented in a system. The results of this work have been published as a Task Force report, a functional model of digital preservation and a user guide to the model.[170] The InterPARES functional model of a digital archive complements the OAIS reference model, having been based, first of all, on archival theory and the requirements of preserving authentic electronic records. The Task Force report concludes: "While the scope of the OAIS model extends far beyond the domain of records, that model could be informed by archival understanding of authenticity."[171]
InterPARES 2 project will continue research into digital preservation, developing prototypes of appraisal and preservation systems, activity models and guidelines for records preservers.[172]

4.3.3 Digital repository management systems

Although the OAIS reference model is explicitly defined as not an implementation guide or a specification for a technological solution, digital repository management systems that claim to be based on the OAIS standard have began to emerge. Some of these are adaptations of commercially available tools, some are being defined and designed by archivists. The explosive growth of digital data in all walks of life have made it clear that new, more powerful and more scaleable systems for managing digital collections are required by a wider circle of institutions than ever before.

ARELDA

The Swiss Federal Archives started developing a solution for archiving data objects from relational and other database environments early 1990s. The scope of the project, named ARELDA ("Archivierung elektronischer Daten und Akten") has over time been extended and the goal of the project is now defined as finding long-term solutions for archiving all digital records in the Federal Archives. The system being developed includes accession, long-term storage and preservation of data and is considering new methods of access for the users of the archives.[173] Since year 2000 ARELDA has a separate project organisation attached to the Federal Archives, because the Archive's funding was not sufficient for an extensive development project. The project is to run at least until 2008 and should result with the Federal Archives being equipped with a solid, long-term and extensible platform for the archiving of digital documents.

Currently the project is implementing a solution for nearline long-term storage with scalable memory capacities of at least 500 Tb. In 2002 the total volume of digital material in the Federal Archives was 6 Tb, with the expected annual growth of 20 Tb.[174] A joint tender with the Swiss National Library recently resulted in the purchase of a new large-scale storage infrastructure. For the Swiss National Library, the equipment will form the basis for storing

[170] cf. http://www.interpares.org/book/index.htm, Ch. 3 and appendices 5-7
[171] http://www.interpares.org/book/index.htm, Ch. 3, p. 14
[172] http://www.interpares.org/ip2/ip2_domain3.cfm
[173] T. Schärli, *Projects and initiatives in Swiss Archives – a bottom-up experience* (2001)
[174] cf. *Archivierung von elektronischen digitalen Daten und Akten der Bundesverwaltung im Schweizerischen Bundesarchiv (ARELDA)* (2001)

electronic publications (e-Helvetica). The repository management system being developed by ARELDA project is based on the functional entities of the OAIS reference model. The multi-stage implementation project should be at the stage of having a repository solution for short-term storage (5 years) in place and a functioning XML-based metadata management system.[175]

e-print digital repository management systems

The setting up of institutional repositories by universities was started by academics with the intention of sharing their academic output for free. The institutional e-print repositories are acquiring more and more functionalities of traditional collections management.[176] Most of these repositories are based on the concept of 'self-archiving' where material can be added to the repository by practically anyone, although in the case of theses collections etc. the repository content is controlled. The software used for managing the digital collection or repository are often available free of charge, not too complicated to install, set up, administer and manage. Due to a wide range of potential depositors with varying degree of archiving skills, the description of materials is kept to a minimum – most repositories use the 15 metadata elements of Dublin Core standard and follow the *Open Archives Initiative Metadata Harvesting Protocol* to enable cross-repository searching.[177]

The digital repository management software can be developed and maintained by a university or institution (e.g., CDSware, eprints.org, i-Tor, etc.); or developed and supported as a co-operation of several universities and institutions (e.g., MyCoRe, DSpace, Fedora, etc.); commercial products are also being developed.[178] Although the main focus of these software tools is on providing easy access to the content of a repository, they will have to begin developing further tools for managing and preserving the collections as their size grows.[179] These functionalities are already being looked into and tested by a few of the software developers.[180]

Digital Object Management System (DOMS)

Several national libraries have developed specialised systems for managing and preserving their digital collections that they call *Digital Object Management Systems* or *Digital Asset Management Systems*. The purpose and structure of these systems is similar to the DIAS system at the Netherlands Koninklijke Bibliotheek that was described above, but there may also be additional functionalities. For example, one of the purposes of the National Library of Australia's *Digital Collections Manager* is managing digitised material and providing access to it.[181] The system being developed there consists of a number of modules and integrates several other systems and repositories that are being maintained by the library.[182] Similar systems are being developed elsewhere, e.g., at the British Library, University of Texas Library, National Library of Scotland, National Library of Finland, etc.

Digital archive management software systems

---

[175] cf. *Archiving of Electronic Digital Data and Records in the Swiss Federal Archives (ARELDA): e-government project ARELDA. Management Summary* (2001) and OFCOM, *6th Report of the Information Society Coordination Group (IISCG) to the Federal Council* (2004)
[176] R. Crow, *SPARC Institutional Repository Checklist and Resource Guide* (2003)
[177] http://www.openarchives.org/OAI/openarchivesprotocol.html
[178] see *A Guide to Institutional Repository Software* (2003); http://www.oaforum.org/resources/tvtools.php
[179] cf. S. Anderson, et al., *Feasibility and Requirements Study on Preservation of E-Prints* (2003)
[180] U. Borghoff, *Vergleich bestehender Archivierungssysteme* (2005)
[181] NLA, *Request for Quotation - Digital Object Management System* (2000); NLA, *Digital Collections Manager Functional Specifications* (2002)
[182] NLA, *Digital Object Storage System* (2001)

Similarly to libraries, archives in many countries have begun defining requirements and tendering for digital archive systems software. Functional requirements for digital archive systems have been developed and tenders announced for example by the U.S. National Archives[183] and the Public Record Office of Victoria in Australia,[184] a tender has been in preparation in Finland. The systems being tendered for are generally based on the OAIS reference model, but their design has been enhanced with functions and activities required in a public archive. Both of the mentioned calls for tender also include elements of software specifications where, in particular the user interface has been defined in detail. The Digital Archive at Victoria has entered the solution development phase now and it is hoped that the system will be functional in 2006. The UK National Archives working together with a company Tessella, developed Digital Archive System that is broadly based on the OAIS standard, although not explicitly designed to be compliant with the standard.[185]

No doubt many other digital archive systems are being developed at this time, and OAIS reference model, as the only internationally accepted standard, acts as the main blueprint for these systems.

4.4 Digital preservation – conclusion

The classical digital preservation theory includes a paradox – preservation of the original record in an authentic, unchanged and usable state is possible only with methods that assume changing the original. All proposed, used and tested methods of digital preservation rely on changing at least one of the components of a digital object (i.e., either the physical, logical or conceptual level). Either the hardware, software or data in the file need to be changed or in some cases all should be transformed during the preservation process. The need for this is arising from the presumption that the future user of the electronic records in the archive needs to see, read, process and use the files in exactly the same ways as the creators of the records did. However, the requirement to maintain the original record and its authenticity and preserve the constituent file unchanged, is contradicting the requirement to keep the file useable on the unknown future technological platforms.[186] Hence, the real task of a digital preservation strategy is to define the requirements of the record as such that they would allow for changes in the lower constituent levels of the record.

The first ten years of serious research into digital preservation issues were largely spent on defining and finding the ideal form for the preservation of digital material (e.g., the storage media with longest useful lifetime, the suitable compression programs, the longest-lasting file formats, etc.). Many of the now classical IT standards were in the 1970s and 1980s only one of many possibilities (e.g., ASCII, SGML, SQL, etc.) and the quest for a single common standard format seemed likely to have a result. With the microcomputer came the proliferation of application software and the endless version updates of software and operating systems that added a new dimension to the digital preservation problem, and gave a clear sign that solutions to the problem are urgently needed. The obvious method of converting files into new, emerging file formats, that was borrowed from computer scientists and used in archives for at least two decades already, began to be criticised because of the changes in digital objects that it caused. For a method for perpetual digital preservation, migration was considered to be too labour and resource-demanding, plus the archivists have no control over the tools they have to

[183] http://www.archives.gov/electronic_records_archives/acquisition_information.html
[184] http://www.prov.vic.gov.au/vers/digitalarchive/development.htm
[185] cf. http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/project_background.pdf
[186] NARA/Electronic Records Archives, *Introduction To Preservation And Access Levels Concepts* (PAL v1.0) (2003), p. 1

use for conversion, and hence the whole development of the method. The emerging alternative strategies – stabilising the digital object into an open standard format, and emulating the current technological platform on a future computer technology, represent an economical, but not ideal solution for today, and the latter a possible future solution. Both approaches attempt to free the archivist from having to convert the whole archive's collection at regular intervals, but it remains to be conclusively proven that either of the proposed solutions can ensure the preservation of records' authenticity and integrity on acceptable levels.

The period of comparisons, debates and criticism has benefited both "camps" – the now already "classical" migration strategy, as well as the emulation strategy and standard formats based preservation methods. Preservation strategies are now more clearly defined – formal definitions, models, criteria and assessment methods, have helped to put these methods into practice. The validity of the emulation strategy has been demonstrated through tests with adequate results, and the method has been developed further by proposing a single, universal emulator or definition of its language. In defining new file formats the need to use the files over long term has began to be considered (cf. standardisation efforts of JPEG and MPEG formats) and some software producers (e.g., Adobe) have made their file format specifications publicly available. The emergence of XML has given a "second wind" to the proponents of standardised file formats, and the archival community is now attempting to lobby the large software producers into developing tools for converting their proprietary file formats into standardised XML. Some of the widely spread file formats have developed their own "archiving versions" (e.g., PDF/A,[187] MPEG-7 variants). Using XML has allowed to create wrappers of metadata and encapsulating digital objects in these has become another method for digital preservation. Migration strategy, as the most widespread and practised method for digital preservation, has reached the phase of looking back and analysing the experiences and results so far. This has led to developing synthesised tools with knowledge about the behaviour of file formats and conversion risks being assembled into registries that everyone who has a digital preservation responsibility could use. Developing information portals and other tools with information on file formats and their conversion instructions is the next logical step forward that will save the time and resources at many institutions who would otherwise have to get involved in costly digital preservation research. It is noteworthy that archives are behind such global 'technology watch' services that these portals and registries will offer and which will be useful for many communities and domains. This will enable to automate decision-making in the migration strategy and, thus, make this approach cheaper and more easy to implement also in smaller archives and other institutions.

Developing new approaches and solutions for the digital preservation problem has, however, not stopped and research into digital preservation issues continues. But the trend of suggesting that a solution found in one discipline or domain could solve everyone's problems with retaining digital material, has been by now been replaced with focussing on the needs of a more limited, specific interest group or a limited implementation area and developing a digital preservation solution for it.

Perhaps the most important statement made in recent years about digital preservation research and practice has come from the InterPARES working group in regard of impossibility of proving the success of digital preservation activities. Methods for digital preservation cannot be separated from the use of digital records – as long as the record has not been retrieved and displayed in a human-readable way, we cannot prove that a record, or any other digital object, has been successfully preserved. Because at the time of creation or archiving it is not possible to foresee the future technology platforms and ways of using digital objects in them, all digital

[187] cf. William LeFurgy, *PDF/A: Developing a File Format for Long-Term Preservation* (2003)

preservation strategies are by default evolutionary, i.e., they must develop together with the technology that they depend on. This perpetual change and development will, however, leave the authenticity of records in an unresolved state since, unlike the developing technology, the authenticity requirements are staying static. The acknowledgement that none of the existing digital preservation strategies satisfy the authenticity criteria of preserving an unaltered record, has put the theorists of digital preservation back to re-thinking the basic concepts, and realisation that solutions to digital preservation problem can only be defined from archive- and records management requirements, rather than from technology and dependency on it.[188] Defining criteria, requirements and significant properties for preserving authentic electronic records is, perhaps, the main trend in current digital preservation research.

Digital preservation strategies derived from technological solutions carry an inherent assumption that the future users of electronic records will want to use and process the electronic record in the same way as its creators were able to do with software available to them. For a great majority of digital material preserved in public archives around the world, the future users will actually only want to read and look at the record, whereas any additional processing of the record would qualify as altering the authentic record, or a new type of service that archives have traditionally not provided to their users. Database records and multimedia materials could be subject to a limited processing, but solutions for users are possible where static subsets could be extracted from the record. Under the recently revived way of thinking about digital preservation, where the focus is on preserving the record (as a conceptual object) and its usability, such re-definitions of what is the purpose of preservation activities, are commendable.

It is clear that there need to be rules for defining criteria for digital preservation and selection of appropriate strategy based on these. A framework for creating such rules has been offered by the OAIS reference model which also provides a functional model of a digital archive. This recent ISO standard proposes a conceptual model, how digital preservation and its decisionmaking are integrated into the functions of a digital archive and connected to both creators and users of records outside the archive. The OAIS reference model is, understandably, broad and general, and has already been tailored and enhanced for specific domains. Two adaptations of the OAIS model have, so far, been published: the research libraries group (RLG) has taken an initiative on studying the relations between digital archive and its users, and developed criteria for trusted digital repositories; the InterPARES project, which has more archival background, has studied the "input" of a digital archive – the authentic electronic records and defined its own model for preserving authentic digital records. Metadata for preserving digital material has also been developed, based on the OAIS informational model. Work on the OAIS reference model and its related topics will no doubt continue for some time yet.

The models for digital archives that these projects have produced are, however, only general level reference models and do not contain specifications for implementing such a system in practice. The real everyday management of digital archives is still reliant on bespoke systems developed by archives themselves or in some few cases, on adaptations of commercial software. Development of integrated digital archive software packages is still only at the stage of defining functional requirements and tendering for partners who could develop the software. Some software developers have begun re-tailoring their existing software solutions and enhancing them to meet the digital archive management needs. These solutions have, as yet, very few tools that could help the archivist with digital preservation activities and documenting these. The number of software packages on the market that use the word

[188] cf. InterPARES, *Preservation Task Force Report* (2002), p. 11

"archive" in their name is impressive, but few of these offer any tools for managing the preservation of the digital collection they manage. It is hoped that further collaboration of archivists and software developers results in better products that meet the needs of professional archivists. This, however, precludes that archivists have made up their mind about which digital preservation strategy they want to implement. As this chapter aimed to show, making these decisions is at the current stage of digital preservation theory, not entirely free of risks.

5. From archive to users

The cost of digital preservation is often connected indirectly or even directly to the methods that are being used or planned to be used for providing access to the preserved records. The closer to the original functionality the access and usage methods have to be, the more strict are the requirements for the preservation process and the more expensive the preservation is likely to be. On the other hand, the entire purpose of preserving something in an archive is that it can be accessed and used in the future. Even though we do not know yet what the future platforms for using electronic records are going to be and what kind of access the future users will require to the records, it is our task to make sure that digital materials in archives carry as little as possible the risk of losing access to them. As a minimum (or indeed, maximum), we can ensure the same level of usability of records and the same functionalities of software as were available to creators of records. Providing this high level of service may prove to be difficult with certain types of complex digital objects, e.g., GIS, CAD/CAM records, databases, etc. Archivists must make compromises between the long-term accessibility of data and their easy usability. What exactly could such levels of service and access provided by a digital archive be, has not been discussed at length, nor agreed yet, although first attempts are being made.

Digital archives that can offer their users flexible access tools already now (e.g., different dissemination file formats, on-line access and analysis tools, etc.) are often not bound by the requirement to preserve authentic digital records (e.g., social science data archives). Archives that keep public electronic records from 1970s and 1980s can at the moment offer access to data in so called low file formats and minimal tools for using them in the archive. Only few modern public electronic records from the "PC-era" are made accessible in archives and rules for their usage methods are only being developed.

There are only a few studies that have looked at what the users actually want to get from a digital archive and how they expect to use a digital archive. These studies have focussed on the needs of a clearly defined user group – schoolchildren (e.g., U.S. Library of Congress, etc.), scientists in a specific field (e.g., as part of the e-Science programme in the UK; the EU funded ARENA project for better use of archaeological finds, etc.), readers of scholarly journals (e.g., JISC in the UK, etc.), . A synthesised analysis of digital archive user needs is absent and research for the current report did not come across projects that are analysing these needs. Some of the issues related to use of digital archives are discussed below.

5.1 Significant properties

The significant properties of records are analysed not only for establishing their authenticity criteria, but also in order to define requirements for their use. The level and type of access archives need to provide to their holdings determines the 'what' and 'how' they need to preserve. Right decisions must be made already at the time of ingest and metadata creation, and all significant properties influencing the usability of records must be retained throughout preservation actions. Research in the area of significant properties of digital objects is directly

related to using the objects in the future (cf. CEDARS and NEDLIB projects, NARA's co-operation with SDSC, etc.).

5.2 Dynamic documents

A good example of problems related to preserving the original functionality of archival records is web documents. These are usually defined as dynamic documents, as their content and volume is often changing in almost real time. In some respect, databases can also be viewed as dynamic objects,[189] but most treatments of dynamic documents are using web documents as examples. How to preserve the content and interactive means of use of these documents for the future users, has not been solved for archives yet.

Archives have by and large stated that web documents can also be public records and need to be transferred to archives. National Archives of Australia and Canada have issued guidelines for treating Internet and intranet documents as corporate records.[190] In the UK, JISC has commissioned a number of works on archiving web documents.[191] Libraries have invested more into studying the issues and solutions for preserving the web (e.g., Australia,[192] Sweden, U.S., the Netherlands, etc.) and have even founded the International Internet Preservation Consortium,[193] but also archives are looking more into this area (e.g., UK National Archives, ERPANet,[194] DLM-Forum, etc.). Using web records in an archive is a relatively new service and publicly documented only at the National Archives in the UK. Archived web pages of the Government Departments can be browsed in the archives' reading room and on-line access is provided using the services of the Internet Archive.[195] The National Archives has also established the UK Web Archiving Consortium, which is to help defining the requirements for web preservation and use of archived pages.[196]

5.3 Automation of finding aids

Archival information systems and catalogues have gone through significant development over the past decade and now can offer on-line access not only to catalogue metadata but also digitised images of archival holdings and electronic records. The XML language has provided new tools for building, integrating and linking catalogues, and the number of inter-archive and international catalogue-mapping exercises and search engines being developed has proliferated. Step-by-step digitised images of and born-digital records are appearing in these catalogue systems.

The National Archives of UK offers its users a number of automated on-line searching tools – the main catalogue of the archive PROCAT, on-line access to archived web pages, on-line access to digitised historical census materials, on-line government databases via NDAD catalogue.[197] The National Archives of Sweden is maintaining the NAD database[198] that contains catalogue information about records in Swedish archives, museums and libraries and that can offer on-line access to electronic records as well. For the time being, JPEG 2000 file

---

189 cf. http://www.erpanet.org/events/2003/bern/index.php
190 http://www.naa.gov.au/recordkeeping/er/web_records/intro.html; http://www.imforumgi.gc.ca/iapproach2_e.html
191 http://www.jisc.ac.uk/index.cfm?name=project_webarchiving
192 http://pandora.nla.gov.au/index.html
193 http://www.netpreserve.org/about/index.php
194 http://www.erpanet.org/events/2003/kerkira/index.php
195 http://www.nationalarchives.gov.uk/preservation/webarchive/default.htm
196 http://www.webarchive.org.uk
197 http://ndad.ulcc.ac.uk/search/
198 http://www.nad.ra.se

format has been chosen for providing access to digital records – all records are converted to this format for access. Possibilities of XML are being explored to offer on-line access to more complex types of electronic records (e.g., databases and registries).[199] The "digital reading room" of Swedish archives SVAR[200] offers both free and fee-based access to digitised images of older archival records. A similar archival information system, called Daisy, is being developed at the Danish National Archives. The Vakka system being developed at the Finnish National Archives is to offer access to digitised archival material as well as catalogue data. The Norwegian National Archives offers direct access to electronic records on a minimal level, but the digitisation and transcription of historical archival material has been developed extensively through the historical digital archive consortium.[201] Estonian National Archives has recently released its archival information management system on Internet, providing on-line access to its catalogue data.

Examples of international and inter-archival co-operation in building on-line search and access tools include the European Visual Archive (EVA)[202] that provides access to digitised photo collections from archives in several countries; Manuscripts and Letters via Integrated Networks in Europe (MALVINE) project[203] that provides details on the nature and location of post-medieval manuscripts held by a wide range of cultural heritage institutions in Europe; the Archives Hub[204] that as a national gateway to descriptions of archives in UK universities and colleges; Access2Archives[205] portal and database that contains catalogues describing archives held throughout England; etc.

Easy access to archives via Internet is dependent on metadata standardisation. ISADG standard has provided a good starting point for this, but more metadata is required for preservation of electronic records and for describing the structure, behaviour and properties of complex digital objects. Standardisation efforts for these metadata are only in progress and the access tools for users of archives are likely to be developing for quite some time yet.

5.4 Digitisation of archives' holdings

Practically all archives around the world are by now involved with digitisation of their collections on a smaller or larger scale. This is mostly driven by the perceived need of improving access to archival materials and disseminating them in new ways. In some projects (e.g., in tropical climate areas) the aim is also 'digitisation for preservation', i.e., creation of archival digital copies that can, in principle, replace the paper original if it deteriorates rapidly. Automation of finding aids and digitisation of collections usually go hand in hand in archives, and the digitised materials are made available via the created on-line catalogue.

Within the library community there exists more experience with scanning textual and microfilmed material: the RLG, OCLC and many other library organisations have published extensively on best practices in digitisation. But archives are quickly catching up in this area and are offering new ideas for disseminating the digitised material. On the one hand, digitisation of paper records is good for the protection of the originals – their use, wear and tear through use, physical retrieval from the repository conditions is reduced to very few occasions and the restoration costs should be minimal. But on the other hand, digitisation adds

[199] *Arkiv för alla – nu och i framtiden* (2002), p. 222
[200] http://www.forskarsalen.ra.se
[201] http://www.digitalarkivet.no
[202] http://192.87.107.12/eva/uk/about.htm
[203] http://www.malvine.org
[204] http://www.archiveshub.ac.uk
[205] http://www.nationalarchives.gov.uk/a2a

new aspects to the digital preservation problems of the archive. Although it is not original, born-digital material, it also requires storage, description, maintenance and digital preservation (e.g., conversion). Managing and preserving the digitised and born-digital records in one archive system can be effective (cf. the Swiss Federal Archives system), but has not always been possible or even been set as a goal (cf. the Norwegian National Archives solution with its historical digital archive). The first collections to be digitised are usually the ones that are most in demand by certain user groups of archives (e.g., family historians) who have clear access and usage needs. Source material for genealogical research, primarily census materials, have been digitised and made available on-line in archives in many countries, and can serve as a source of additional revenue for archives (cf. the National Archives in the UK, the ARKIS system in Sweden, etc.).

5.5 Access to digital archives – conclusions

Providing access to digital archives has so far been very seldom the subject of research projects, and a common international best practice has not emerged yet. Modern digital records have reached national archives only recently and in some countries are still closed for users – these archives are only beginning to define the ways they want to make their digital records accessible. Archives with older digital public records (from 1970s and 1980s) that are mostly statistical and numeric material held in databases and registries, offer minimal access tools to their users, although several social science data archives have developed also on-line analysis tools for their collections. Compared with a decade ago, archives and libraries now have considerably more experience with and means for access to digital material, both for general use and as a fee-based service. These services have first been developed for user groups with clearly defined usage requirements, which demonstrates the need for further research into archives' users needs and expectations in order to define better tools and systems for access to digital archives.[206]

The so called social science data archives that primarily collect data resources created by social science and humanities scholars, have developed more advanced tools for accessing and using their digital collections. Other archives have largely ignored the long experience of data archives with digital preservation and tools developed for providing access to digital archival material, although for instance the XML-based interactive analysis tools and flexible user service could be a very useful example to many archives.

When the purpose of digital preservation is explained as the preservation of integrity and usability of digital records, then the ability to use the record is one of the criteria by which the success of preservation activities are assessed. However, this assessment can only be carried out in the future, on a future computer technology platform. This built-in paradox will be part of any digital preservation strategy – whether the strategy successfully ensures the usability of the record can only be ascertained when the record is used some time in the future, not at the time of selecting on or another preservation strategy. Research into significant properties of digital records is helping to improve our understanding of changes our preservation activities may inflict on the records, and help us to make better informed decisions in the course of digital preservation. But the answers cannot be all found within the records alone, we must also turn to the users of these records and study their requirements. The OAIS reference model has made the first step in connecting the end-user (designated community in OAIS-talk) and the description of the digital objects in archive (cf. ch. 3.2.3 above). The OAIS has approached

[206] cf. H. Kemoni, et al., *Obstacles to utilization of information held by archival institutions: a review of literature* (2003)

this issue on a general level, more detailed analysis and results are eagerly awaited by the archival community.

6. Conclusion – a look into the future

Nearly ten years ago, in 1996, a Task Force commissioned by the Commission on Preservation and Access and the Research Libraries Group in the U.S., issued a report on preserving digital information that included a recommended set of research issues around long-term preservation:[207]

- Define requirements and standards for describing and managing digital information.
- Define migration paths for digital preservation of culturally valuable digital information
- Research to define acceptable levels of information loss during migration and identify a set of minimal record attributes, which if not preserved, would make investments in preservation pointless.
- Research on migration should examine the physical attributes of digital objects.
- Research on degrees of functionality that are needed or desirable when preserving digital information.
- Address the requirements for retaining the relationships among digital objects.
- Research and development of cost models for the various approaches to preservation.

Computer technology has been developing rapidly within the past ten years and new issues have been raised for digital long-term preservation, but some of these issues are still with us. In 2002, a group of computer scientists, information scientists, archivists, digital library experts, and government program managers met to examine the prospects for advancing computer and information technology research through a research program that addresses the unique challenges of long-term digital preservation. The Workshop on Research Challenges in Digital Archiving and Long-term Preservation published a report with a research agenda around four main themes: 1) technical architectures for archival repositories; 2) attributes of archival collections; 3) digital archiving tools and technologies; and 4) organisational, economic, and policy issues.[208]
In 2003, an NSF-DELOS Working Group on Digital Archiving and Preservation summarised these topics and published a further elaborated research agenda for digital archiving and preservation that included the following topics:[209]

- Elaboration of existing repository models leading to technical specifications and standards that can be used to build persistent archives.
- Establishing registries of digital formats that 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.
- 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.
- Research focussing on context sensitivity, risk awareness and proper preservation behaviour.
- 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.

---

[207] see CPA/RLG, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information* (1996); M. Hedstrom, *Research Issues in Migration and Long-Term Preservation* (1997)
[208] NSF/LoC, *It's about Time. Research Challenges in Digital Archiving and Long-Term Preservation* (2003)
[209] NSF-DELOS, *Invest to Save* (2003)

- Preservation processes that can be automated need to be identified and mechanisms for automating them developed, because human intervention at each stage of the preservation process is not economically viable.
- Tools and methods for detecting trustworthiness and information quality.
- Scalability of both preservation processes to large collections, and developing inexpensive preservation tools and technologies that individuals without extensive archival or IT skills can readily use.
- How can the creation and authoring of metadata be automated and how can we ensure long-term metadata viability?

These lists are, to an extent, overlapping, which demonstrates that not all satisfactory answers have been found as yet. Some observers have hastily concluded that nothing radically new is happening in the digital preservation research, since most areas have been under discussion for at least a decade already. A real advance has been, nevertheless, made in testing the results of this research in practice, implementing systems that are based on new theories and getting the hands-on experience with the utility of methods developed during the pioneering stages of digital preservation research. Some research areas are likely to continue to be of interest for a longer period. The list following below is by far not complete — it emerges from the materials collected for this report, which is limited in its geographical and topical scope.

6.1 Cost models for digital archiving

For quite a number of years it has been difficult to estimate the real cost of preserving records digitally. Both in real terms and as compared to preserving paper records. The costs would also include the expenses for setting up a digital repository or a digital archive service, periodic archiving and use of electronic records.

The real cost of archiving and preservation is depending on the nature and authenticity requirements of materials, and the level of access that is provided, or promised, to the archived resources. Literature on this subject has become more extensive in recent years, also in the context of institutional repositories. But it should be possible to advance the cost modelling further than just a few file formats and case studies of individual institutions that have been published so far. Better costing methods are also needed in order to convince the fundmakers that they receive a good return on their investments into digital preservation services.

6.2 Automation and scalability of digital archives

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. As the volume of digital material reaching the archives is growing (or even multiplying) the existing methods and practices of treating this material are put to a very harsh test indeed. Can repositories be designed to handle collections with billions of digital entities in dozens of different formats? Will the current validation tools in archives be fast enough to treat terabytes of acquired data at a time? Can a large collection of records be migrated within a feasible timeframe? Can some or most of the digital archive's work processes be automated?

NARA in the U.S., in co-operation with SDSC, has used files on the Internet to test the processing of bulk records and explored the technical architecture that is capable of handling a significant volume of data and records (e.g., a million e-mails). Still, finding and retrieving a single item from such a 'corpus of records' remains an issue that is not only technical but also intellectual or even logistical.

Digital preservation and research for it has largely been the "prerogative" of large institutions. Their "problem" with the digital preservation is usually more acute and complex, and they often have better access to funding sources. The expensive digital archive systems being

developed, built, tendered and acquired by larger archival institutions and repositories, do not solve the problem for smaller institutions who may have equally valuable, but smaller scale digital collections. Solution to the digital preservation problem cannot be found in large-scale technology alone, it has to be flexible and scalable. Scalable architectures for digital repositories that would suit both small and large collections of records, yet maintain the same level of processing and retrieval capabilities are urgently required. The rapid growth of emerging institutional repositories is a good example of an area where libraries and archives have missed an opportunity of offering small-scale and flexible solutions and services, where there has been a real market need for them.

6.3 Benchmarking

More standards — international, national and discipline-specific — are being developed for records and archives management than ever before. Standards make it easier to compare the practices between archives and to a benchmark on international best practice that the standards represent. The rising and levelling quality of services and practices in archives will help the preservation of digital material as a whole and also in each individual case. Certification of repositories and developing criteria for trust are one example how an international standard (OAIS reference model) can serve as a basis for a real benchmarking exercise. Another example of standardisation of knowledge about the behaviour of digital material that has accumulated in archives are the file format information portals and registries. When completed, these benchmarks and tools for migration planning will help all interest groups involved in digital preservation.

6.4 Risk analysis

Risk analysis has become a popular topic in records management and archives, as well as in general management theory. With an increasing number of standards becoming available to archives, archivists and records managers, it becomes easier to perform a systematic 'reality check' of the current practices and strategies against the standards and to assess the risks arising from deviations from the standard practices.

In some respect the digital preservation itself is risk management – with no absolutely guaranteed methods available for preserving the digital materials for long-term, archivists are trying to minimise the risk of lost access to the records and attempting to maintain its original usability. The risk of loosing access to a digital object due to technology obsolescence has been the cornerstone of understanding the digital preservation and defining strategies for it. Assessing these risks has, however, not been based on a very rigorous methodology.

Risk analysis of file formats has been part of the digital preservation research for a while now. New risk assessment tools, checklists and manuals are now appearing for other topics as well, for example, risks in preservation management, electronic records risk assessment, assessment of organisational capacity to manage records, etc., etc. Research into possible impacts of implementing something new (e.g., a new practice or strategy) should, however, become more commonplace.

6.5 Preservation of web and dynamic records

Web archiving on a significant scale has so far only been accomplished in libraries and projects with specific funding (e.g., the Internet Archive). Archives have so far limited themselves with a few guidelines (e.g., National Archives of Australia and Canada). Serious treatment of archival approaches to capturing, retaining and preserving the accessibility to dynamic compound documents is yet to be developed. With the current drive in businesses towards Content Management Systems, Knowledge Management Systems and Enterprise Portals, the presentation, exchange and gathering of information in the form of dynamic documents is clearly on increase. Practical solutions to archiving this type of dynamic material, without digressing into the discussion of what exactly might constitute a record in this context, are necessary in the nearest future. A few guidelines for treating intranet

documents as records exist (UK, Canada, Australia), but these are mostly based on office systems on client-server architecture, rather than HTML, XML, etc.

6.6 Metadata

Research and discussions over metadata will probably never cease, as long as there are new user groups and new methods for preservation of digital material. Digital preservation as a method relies heavily on detailed, record and file level metadata which archives have traditionally not created nor managed. Compatibility of metadata required in a digital archive with metadata created with records in the records management phase, compatibility with the archival description and finding aids, are issues that have received surprisingly little attention in research projects. Interoperability of metadata, using semantic networks, knowledge maps and other modern tools that have become more accessible to memory institutions with the use of computer technology, are only beginning to be explored by archivists.

Standardisation of metadata for the purpose of interaction and exchange with other institutions (e.g., libraries, museums, various repositories) has started in archives, but there is a long way to go yet. The digital preservation metadata recommendations made within the framework of the OAIS reference model has made a significant step towards offering a common set of metadata for all institutions involved in digital preservation. But an archival treatment of these metadata is yet to be undertaken.

6.7 Records created with open-source software

Open Source Software is software with an openly published source code, it is usually available free or at a small charge, and is often developed through voluntary efforts. First guidelines for using open source software in public administration have appeared in Germany, Italy, Netherlands, the UK and elsewhere. These guidelines do not include description and archiving of records created with open source software. Being cheap and easily customisable to individual needs makes such software attractive to users, but it may become difficult for archivists to keep track of the many "mutations" and changes in such software, and to assess their impact on the preservation of records created with the software.

Archivists need to explore the impact of open source operating systems and software to long-term preservation of records created with them; archives have to develop or adapt metadata standards for describing records, and their properties, created with open source software.

Failure to address digital preservation problems is analogous to squandering potential professional, personal and economic gains, contributing to cultural and intellectual poverty and resulting in exorbitant costs for recovery.[210] Archivists are compelled to meet the research challenge to resolve the apparent conflict between the creation context and the future use context of electronic records to facilitate digital information preservation. Successful planning for the digital future depends upon continuing research and the development and testing of strategies.

On the basic technical level, digital preservation combines two processes: maintaining an accurate byte stream, and maintaining the ability to retrieve and recreate the byte stream on current or future technology. However, repercussions of these procedures and a plethora of other issues that are connected to them have changed profoundly the archivists' thinking on what it is they are doing, what they should be doing, where their responsibilities lie, how they should be funding their activities and what they should research. The latter has been the focus of the current report — the aim has been to describe the current research into preservation of digital information, predominantly in the archives around the world, but also making references to the work in other involved disciplines and stakeholders.

[210] S.S. Chen, *The Paradox of Digital Preservation* (2001), p. 6

The need to preserve digital archives has forces the archivists to get more intensively involved with records management, metadata creation and management, systems for retrieval, access and use of digital materials, and many other issues. The position of the archive, as it were between the creator of the records and their user, dictates that archives should be involved in the creation of records, control of their management processes, and design of new access services to records. And on top of all that the archive is still responsible for preserving authentic, reliable and complete electronic records for the long term.

The report has, in a way, confirmed the recent admission by one of the key figures in digital preservation research: "despite debates in the digital preservation community about the best method for ensuring longevity of digital materials, most recent progress is the result of a focus on particular aspects of the problem and attempts to find solutions to smaller pieces of the puzzle".[211] Digital preservation strategies have been discussed for more than ten years and it is only natural that this discussion continues, with results from practical testing and implementation being taken into account. Successful implementation of an archival preservation system and using it in day-to-day work of a digital archive is the ultimate measure of usefulness of a preservation strategy.

Working on different strategies and models for digital preservation is helping to understand how different approaches contribute to the common framework of solving the same problem and what are the needs of different stakeholders. However, a nearly ten-year old statement that "the Americans tend to put their faith in software and systems to solve electronic records problems, while the Europeans feel that regulation and legislation have an important role to play",[212] has not been proved entirely wrong by the projects and developments reviewed in this report. In European countries a considerable effort goes into developing regulations, recommendations and guidelines, while the technological thinking lags perhaps somewhat behind when compared with Northern American developments. Some of this may be related to funding of archival research projects which has been difficult to achieve in smaller European countries, although in the opinion of many the examples of Switzerland and Sweden represent the case to prove exactly the opposite. Digital preservation is an imperative also for smaller archives and they, too, need to find a good balance between developing technical solutions and producing regulations and guidelines, while keeping up-to-date with the theory at the same time — this is a "trick" that no research project has provided a guideline for, yet.

An important trend in the digital preservation research in archives in recent years has been the closer collaboration with other interest groups and stakeholders. The understanding that other memory institutions need and offer similar services with digital materials, as archives do, has led to wider co-operation, joint projects and generally better use of resources available for research and developing new services. Archives and libraries are co-operating for example in defining research agenda for long-term digital preservation, certification of digital repositories, defining interoperability rules for metadata, file format research, etc. Different stakeholders are analysing the digital preservation strategies and models for managing digital collections, which is helping to make it clearer for everyone what the different approaches to the same problem are, what are the different requirements of various stakeholders and how the solutions can be achieved jointly.

The conclusion by archivists that a digital record cannot be preserved in the traditional meaning of this word — it is only possible to retain the ability to reproduce the record — has, perhaps, been the most important step in the theory of digital preservation in recent years. This means that digital preservation is not only about physically preserving the byte-stream and the

[211] M. Hedstrom, *Digital Preservation: Problems and Prospects* (2001), p. 3
[212] cited here from G. Mackenzie, *Searching for Solutions: Electronic Records Problems Worldwide*, 2000, p. 59

software necessary for using it. The archive no longer is a "neutral communication channel" that carries information into the future without changing it — the archive has to make decisions how to make the necessary changes to the information it preserves, what alterations are allowed, what changes are good for the preservation of accessibility to the information. Taking the preservation of integrity of the record as the basis, the archive may choose different preservation strategies and need not be tied to one or another approach.

The questions that need solving are by far not easy to answer — as one librarian noted "suddenly books seem very appealing in their simplicity – the only delivery system required is the ability to read". Archivists need to think of and produce not only of working delivery systems, but also creation, management, preservation and description systems for electronic records. To achieve this, it might be helpful to adopt a dynamic way of thinking — what has recently been called 'the archive as an ecosystem'.[213] Because electronic media, software applications and computer hardware will all continue to change and develop at a rapid rate and because policies must be developed to address that reality, the archive must change as well. The very notion of a permanent archive that is fixed for long-term, may have to give way to an "ecological" preservation system that (perhaps paradoxically) is in a state of constant change. Information grows, lives and dies, as do delivery systems and the task of keeping information alive requires frequent adaptations to and perpetual evolution of the archival system. Rather than making temporary plans in the hopes of a future permanent solution, it might be advantageous to think of all information preservation as an evolving, ever-changing ecosystem. What is necessary then, is a permanent strategy for handling perpetual change. Joining the forces with other stakeholders who face the same problems of maintaining access to digital resources over long-term, could only benefit archives.

[213] cf. J. Martin, D. Coleman, *Change the metaphor: The Archive as an Ecosystem* (2002)

7. Bibliography

"A Guide to Institutional Repository Software" (2003) http://www.soros.org/openaccess/pdf/OSI_Guide_to_Institutional_Repository_Software_v1.pdf

AHDS, "Managing Digital Collections. AHDS Policies, Standards and Practices" (2001) http://ahds.ac.uk/collections.pdf

S. Anderson, H. James, S. Pinfield, R. Ruusalepp, "Feasibility and Requirements Study on Preservation of E-Prints" (2003) http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

"Archivierung von elektronischen digitalen Daten und Akten der Bundesverwaltung im Schweizerischen Bundesarchiv (ARELDA)" (2001) http://www.bar.admin.ch/webserver-static/docs/d/arelda_expose_0301_d.pdf

"Arkiv för alla – nu och i framtiden", SOU 2002:78, Stockholm (2002)

"Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources", RLG/OCLC (2001) http://www.rlg.org/longterm/attributes01.pdf

David Bearman, "Reality and Chimeras in the Preservation of Electronic Records" // *D-Lib Magazine*, vol. 5 (1999), no. 4 http://www.dlib.org/dlib/april99/bearman/04bearman.html

John Bennett, "A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material", *British Library Research and Innovation Report*, no. 50, British Library Research and Innovation Centre (1997) www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html

U. Borghoff, "Vergleich bestehender Archivierungssysteme", Nestor-materialien, no 3 (2005) http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_03.pdf

CEDARS, "Cedars Guide to the Distributed Digital Archiving Prototype" (2002) http://www.leeds.ac.uk/cedars/guideto/cdap/guidetocdap.pdf

J. Coleman, D. Willis, "SGML as a Framework for Digital Preservation and Access", CLIR (1997)

CPA/RLG, "Preserving Digital Information. Report of the Task Force on Archiving of Digital Information" (1996) http://www.rlg.org/en/page.php?Page_ID=20442

Raym Crow, "SPARC Institutional Repository Checklist & Resource Guide" (2003) http://www.arl.org/sparc/IR/IR_Guide.html

Marcel van Dijk, "It Always Hurts the First Time: Experiences with transferred electronic records" // *Cultivate Interactive*, no. 9 (2003) http://www.cultivate-int.org/issue9/amsterdammro

Digital Preservation Testbed, "Emulation: Context and Current Status" (2003) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/White_paper_emulation_UK.pdf

Digital Preservation Testbed, "Migration: Context and Current Status" (2001) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf

Digital Preservation Testbed, "XML and Digital Preservation" (2002) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf

Charles Dollar, "Authentic Electronic Records: Strategies for Long-Term Access" (1999)

Luciana Duranti, "'From Here to Eternity:' Concepts and Principles for the Management of Electronic Records", paper given at the International Conference of the Association of Catalan Archivists "Los documentos electronicos" (1999)

Federal Archives of Switzerland, “Archiving of Electronic Digital Data and Records in the Swiss Federal Archives (ARELDA): e-Government Project ARELDA. Management Summary” (2001) http://www.bar.admin.ch/webserver-static/docs/e/arelda_expose_0301_e.pdf

Bendert Feenstra, “Standards for a DSEP Standards for the Implementation of a Deposit System for Electronic Publications (DSEP)”, *NEDLIB Report Series*, no. 4 (2000) http://www.kb.nl/coop/nedlib/results/NEDLIBstandards.pdf

Steve Gilheany, “Expected Usable Lifetime of Different Electronic Formats” (2000) http://www.archivebuilders.com/whitepapers/22023v006p.pdf

Stewart Granger, “Emulation as a Digital Preservation Strategy” // *D-Lib Magazine*, vol. 6 (2000), no. 10 http://www.dlib.org/dlib/october00/granger/10granger.html

Stewart Granger, “Digital Preservation & Emulation: from theory to practice”, paper given at the *ICHIM 2001* (2001) http://dspace.dial.pipex.com/stewartg/sgichim.htm

S. Granger, K. Russell, E. Weinberger, “Cost elements of digital preservation” (2000) http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc

Jon Atle Haugen, “Arkivenes plass i den informationsteknologiske utviklingen”, *Rapporter til 19. Nordiske Arkivdage år 2000* (2000), pp. 103-146

Margaret Hedstrom, “Research Issues in Migration and Long-Term Preservation” // *Archives and Museum Informatics*, vol. 11 (1997), no. 3-4

Margaret Hedstrom, “Digital Preservation: Problems and Prospects”, paper presented at the *DL20, University of Library and Information Science* (2001) http://www.si.umich.edu/CAMILEON/camileon%20Presentations/margaretpresentation.pdf

M. Hedstrom, C. Lampe, "Emulation vs. Migration: Do Users Care?" // *RLG DigiNews*, vol. 5 (2001), no. 6 http://www.rlg.org/preserv/diginews/diginews5-6.html#feature1

A. Heminger, S. Robertson, “Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents”, *Proceedings of the Sixth Delos Workshop ‘Preservation of Digital Information’* (1998) http://www.ercim.org/publication/ws-proceedings/DELOS6/rosetta.pdf

David Holdsworth, “Architecture of CEDARS demonstrator” (2001) http://www.leeds.ac.uk/cedars/archive/architecture.html

David Holdsworth, “C-ing ahead for digital longevity” (2001) http://129.11.152.25/CAMiLEON/dh/cingahd.html

D. Holdsworth, D. Sergeant, “A Blueprint for Representation Information in the OAIS Model” (2000) http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF

D. Holdsworth, P. Wheatley, “Emulation, Preservation and Abstraction” // *RLG DigiNews*, vol. 5 (2001), no. 4 http://www.rlg.ac.uk/preserv/diginews/diginews5-4.html#feature2

InterPARES Authenticity Task Force Report (2002) http://www.interpares.org/display_file.cfm?doc=ip1_atf_report.pdf

InterPARES, “Preservation Task Force Report” (2002) http://www.interpares.org/display_file.cfm?doc=ip1_ptf_report.pdf

ISO 14721:2003 “Space data and information transfer systems – Open archival information system – Reference model” http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3=

H. Kemoni, J. Wamukoya, J. Kiplang’at, “Obstacles to utilization of information held by archival institutions: a review of literature” // *Records Management Journal*, vol. 13 (2003), no. 1, pp. 38-42

G. Lawrence, W. Kehoe, O. Rieger, W. Walters, A. Kenney, “Risk Management of Digital Information: A File Format Investigation”, CLIR (2000) www.clir.org/pubs/abstract/pub93abst.html

K-H. Lee, O. Slattery, R. Lu, X. Tang, V. McCrary, “The State of the Art and Practice in Digital Preservation” // *Journal of Research of the National Institute of Standards and Technology*, vol. 107 (2002), no. 1, pp. 93-106

William LeFurgy, “Levels of Service for Digital Repositories” // *D-Lib Magazine*, vol. 8 (2002), no. 5 http://www.dlib.org/dlib/may02/lefurgy/05lefurgy.html

Raymond Lorie, “The Long Term Preservation of Digital Information” (2000) http://www.si.umich.edu/CAMILEON/Emulation%20papers%20and%20publications/Lorie.pdf

Raymond Lorie, “A Project on Preservation of Digital Data” // *RLG DigiNews*, vol. 5 (2001), no. 3 http://www.rlg.org/preserv/diginews/diginews5-3.html

Raymond Lorie, “The UVC: a Method for Preserving Digital Documents - Proof of Concept” (2002) http://www-5.ibm.com/nl/dias/resource/uvc.pdf

B. Ludäscher, R. Marciano, R. Moore, “Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives” // *ACM SIGMOD Record*, vol. 30 (2001), no. 3, pp. 54-63 http://www.sdsc.edu/~ludaesch/Paper/kba.pdf

B. Ludäscher, R. Marciano, R. Moore, “Towards Self-Validating Knowledge-Based Archives”, *11th Workshop on Research Issues in Data Engineering (RIDE), Heidelberg, Germany, IEEE Computer Society* (2001) http://www.npaci.edu/DICE/Pubs

George Mackenzie, “Searching for Solutions: Electronic Records Problems Worldwide” // *Managing Information*, vol. 7 (2000), no. 6, pp. 59-65

J. Martin, D. Coleman, “Change the metaphor: The Archive as an Ecosystem” // *The Journal of Electronic Publishing*, vol. 7 (2002), no. 3 http://www.press.umich.edu/jep/07-03/martin.html

Nancy McGovern, “Preservation storage and processing considerations”, *Cornell University Electronic Student Records Systems Project Report* (2000) http://rmc.library.cornell.edu/online/studentRecords/reportB.htm

P. Mellor, P. Wheatley, D. Sargeant, “Migration on Request: A Practical Technique for Preservation” (2002) http://www.si.umich.edu/CAMILEON/reports/migreq.pdf

Reagan Moore, “Knowledge-Based Persistent Archives”, paper given at the *International Archivists' Workshop* (2001) http://www.sdsc.edu/NARA/Publications/NARA_Archivists.ppt

Reagan Moore, “The San Diego Project: Persistent Objects”, *Proceedings of the Workshop on XML as a Preservation Language* (2002) http://www.sdsc.edu/NARA/Publications/persistent-objects.doc

Reagan Moore, “Preservation of Data”, *SDSC Technical Report* (2003) http://www.sdsc.edu/NARA/Publications/data-preservation.doc

R. Moore, C. Baru, A. Rajasekar, B. Ludäscher, R. Marciano, M. Wan, W. Schroeder, A. Gupta, “Collection-Based Persistent Digital Archives” Part 1 & 2 // *D-Lib Magazine*, vol. 6 (2000), no. 3, 4 http://www.dlib.org/dlib/march00/moore/03moore-pt1.html, http://www.dlib.org/dlib/april00/moore/04moore-pt2.html

NARA/Electronic Records Archives, “Introduction to Preservation and Access Levels Concepts” (2003) http://www.archives.gov/electronic_records_archives/pdf/policies_templates_requirements.pdf

National Archives of New Zealand, “What is a Corporate Record?” (2003) http://www.archives.govt.nz/continuum/dls/pdfs/f1-corporate-record.pdf

National Library of Australia, “Digital Collections Manager Functional Specifications” (2002) http://www.nla.gov.au/dsp/doms/dcm.html

National Library of Australia, “Digital Object Storage System” (2001) http://www.nla.gov.au/dsp/doss/doss.doc

National Library of Australia, “Request for Quotation - Digital Object Management System” (2000) http://www.nla.gov.au/dsp/doms/

NEDLIB, “NEDLIB Contribution to the review of the OAIS Reference Model” (2000) http://www.kb.nl/coop/nedlib/results/OAISreviewbyNEDLIB.html

NSF-DELOS Working Group on Digital Archiving and Preservation, “Invest to Save” (2003) http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf

NSF-LoC, “It’s about Time. Research Challenges in Digital Archiving and Long-Term Preservation. Final Report of the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation. April 12-13, 2002” (2003) http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf

John Mark Ockerbloom, “Archiving and Preserving PDF Files” // *RLG DigiNews*, vol. 5 (2001), no. 1 http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2

OCLC/RLG, “Trusted Digital Repositories: Attributes and Responsibilities” (2002) http://www.rlg.org/longterm/repositories.pdf

OFCOM, “6th Report of the Information Society Coordination Group (IISCG) to the Federal Council” (2004) http://www.isps.ch/site/attachdb/show.asp?id_attach=857

Raimo Pohjola, “Implications of electronic signatures – the situation in Finland. The act on electronic service in the administration and the act on electronic signature” // *INSAR Supplement VII* “Proceedings of the DLM-Forum 2002. @ccess and preservation of electronic information: best practices and solutions. Barcelona, 6–8 May 2002” (2002), pp. 490-494

PROV, “Management of Electronic Records”, PROS 99/007 (Version 2) http://www.prov.vic.gov.au/vers/standards/pros9907vers2/default.htm

A. Rajasekar, R. Marciano, R. Moore, “Collection-Based Persistent Archives”, paper given at the *16th IEEE Symposium on Mass Storage Systems* (1999) http://www.npaci.edu/DICE/Pubs/persistent_archives.ps

RLG/OCLC, “Trusted Digital Repositories: Attributes and Responsibilities” (2002) http://www.rlg.org/en/pdfs/repositories.pdf

Jeff Rothenberg, “Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation”, CLIR & ECPA (1999) http://www.clir.org/pubs/reports/rothenberg/contents.html

Jeff Rothenberg, “An Experiment in Using Emulation to Preserve Digital Publications”, Koninklijke Bibliotheek (2000) http://www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf

J. Rothenberg, T. Bikson, “Preservation: Carrying Authentic, Understandable and Usable Digital Records Through Time”, RAND-Europe (1999) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

K. Russell, D. Sergeant, “The Cedars Project: Implementing a Model for Distributed Archives” (1999) http://www.rlg.ac.uk/preserv/diginews/diginews3-3.html

Thomas Schärli, “Projects and initiatives in Swiss Archives — a bottom-up experience”, *Il Mondo degli Archivi. Special issue: VI European Congress on Archives. Firenze 2001.*

Abstracts (2001), pp. 50-53
http://www.anai.org/Conferenza%20europea/abstracts/3105_sessmatt.htm

Thom Shepard, "Universal Preservation Format (UPF): Conceptual Framework" // *RLG DigiNews*, vol. 2 (1998), no. 6 http://www.rlg.org/preserv/diginews/diginews2-6.html#upf

J. Slats, R. Verdegem, "Practical experiences of the Dutch Digital Preservation Testbed" // *VINE*, vol. 34 (2004), no. 2, issue 135, pp. 56-65
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Article_in_VINE_2004.pdf

"Survey and assessment of sources of information on file formats and software documentation", The Representation and Rendering Project, University of Leeds (2003)
http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf

Kenneth Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years", in: CLIR, "The State of Digital Preservation: An International Perspective" (2002), pp. 4-31 http://www.clir.org/pubs/reports/pub107/pub107.pdf

K. Thibodeau, R. Marciano, "Building the Archives of the Future. NARA's Electronic Records Archives Program", paper given at the *Management of Electronic Records*, October 2-4, 2000, Chicago (2000) http://www.npaci.edu/DICE/Pubs/mer2000.ppt

Steve Thomas, "File Formats for Electronic Text" (2002)
http://www.library.adelaide.edu.au/~sthomas/papers/etext-formats.html

Dan Tørning, "Handling of the electronic records issue and cooperation with public administration. Experience of the Danish National Archives" // *INSAR Supplement IV*: "Proceedings of the DLM-Forum on electronic records. Brussels, 18-20 December 1996" (1997), pp. 85-89

US InterPARES Project, "Findings on the Preservation of Authentic Electronic Records" (2002) http://www.gseis.ucla.edu/us-interpares/pdf/InterPARESInterpreted.pdf

Titia van der Werf, "Long-term Preservation of Electronic Publications. The NEDLIB project" // *D-Lib Magazine*, vol. 5 (1999), no. 9
http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html

Titia van der Werf, "The Deposit System for Electronic Publications: a Process Model", *NEDLIB Report Series*, no. 6 (2000)
http://www.kb.nl/coop/nedlib/results/DSEPprocessmodel.pdf

Titia van der Werf, "Experience of the National Library of the Netherlands", *The State of Digital Preservation: An International Perspective*, CLIR (2002)
http://www.clir.org/pubs/reports/pub107/vanderwerf.html

M. Wettengel, A. Engel, "Disposition and archiving of electronic records: Concepts for the Information Network Berlin/Bonn" // *INSAR Supplement IV*: "Proceedings of the DLM-Forum on electronic records. European citizens and electronic information: the memory of the Information Society. Brussels, 18-19 October 1999" (2000), pp. 102-109

Paul Wheatley, "Migration – a CAMiLEON discussion paper" // *Ariadne*, no. 29 (2001)
http://www.ariadne.ac.uk/issue29/camileon/

J. Wing, J. Ockerbloom, "Respectful Type Converters For Mutable Types" (1999) http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/compose/ftp/pdf/paper.pdf

Appendix 1. Matrix of digital preservation research projects and research topics

Table on the next page presents an overview of digital preservation research topics that are or have been on the agendas of research projects around the world. The selection of both topics and research projects/institutions is based on the current report and published results. Threfore, the table does not include all digital preservation research that is being conducted internationally. The matrix is intended only as an index and tool for using the report that precedes it.

The projects and institutions listed in the matrix are the following:

CAMiLEON (Creative Archiving at Michigan and Leeds: Emulating the Old on the New) (UK, US)
http://www.si.umich.edu/CAMILEON/

Cedars (CURL Exemplars in Digital ArchiveS) (UK)
http://www.leeds.ac.uk/cedars/

DAVID (Digitale Archivering in Vlaamse Instellingen en Diensten) (Belgium)
http://www.antwerpen.be/david/website/default.htm

DCC (Digital Curation Centre) (UK)
http://www.dcc.ac.uk/

DPC (Digital Preservation Coalition) (UK)
http://www.dcc.ac.uk/

Digital Preservation Testbed (the Netherlands)
http://www.digitaleduurzaamheid.nl/home.cfm

DLM-Forum
http://europa.eu.int/historical_archives/dlm_forum/

ERPANET (Electronic Resource Preservation and Access Network)
http://www.erpanet.org/

InterPARES (International Research on Permanent Authentic Records in Electronic Systems)
http://www.interpares.org/

KBSt (Koordinierungs- und Beratungsstelle der Bundesregierung für Informationstechnik in der Bundesverwaltung) (Germany)
http://www.kbst.bund.de/

KB/IBM – Koninklijke Bibliotheek (the Netherlands)
http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study.html

National Archives of Australia
http://www.naa.gov.au/

National Archives of Canada
http://www.collectionscanada.ca

National Archives of Denmark
http://www.sa.dk/

National Archives of Norway
http://www.riksarkivet.no/

National Archives of United States (NARA)
http://www.archives.gov/

National Archives of Sweden

http://www.ra.se/ra/index.html

National Library of Australia
http://www.nla.gov.au/

NEDLIB
http://www.kb.nl/coop/nedlib/

Pennsylvania University Library
http://tom.library.upenn.edu/

RLG (Research Libraries Group)
http://www.rlg.org/

SDSC (San Diego Supercomputing Center) (U.S.)
http://www.sdsc.edu/NARA/Publications.html

Swiss Federal Archives
http://www.bar.admin.ch/bar/engine/Home

TNA (The UK National Archives)
http://www.nationalarchives.gov.uk/preservation/

PROV (Public Records Office Victoria) (Australia)
http://www.prov.vic.gov.au/

Further information about research projects and institutions involved with preservation of digital content can be found for example on the ERPANET project website (www.erpanet.org), under “erpaAssessments”.

Digital preservation research projects and topics

Project **Research topic / Subject**	CAMi-LEON	CED-ARS	DAVID	DCC (UK)	DPC (UK)	Digital Preser-vation Testbed (Nether-lands)	DLM-Forum	ERPA-NET	EU IDABC Prog-ramme	Inter-PARES	KBSt et al. (Ger-many)	KB/IBM (Nether-lands)	NA of Aus-tralia	NA of Can-ada	NA of Den-mark	NA of Nor-way	NA of Swe-den	NA of U.S.	NL of Aus-tralia	NED-LIB	Penn-syl-vania UL	RLG	SDSC	Swiss Arch-ives	TNA (UK)	PROV (Aust-ralia)
Guidelines for managing digital records		+	+		+	+	+	+					+					+							+	
ERMS, their design, functional requirements and issues									+		+		+	+	+	+								+	+	+
Records management metadata													+	+										+	+	+
Digital signatures, PKI and encryption			+								+			+				+								
Appraisal							+	+		+																
Transfer of electronic records to archives								+					+			+	+	+								
Emulation	+	+				+						+								+						
Migration	+	+	+			+						+									+		+		+	
Encapsulation																										+
Persistent archives																		+					+			
Storage media issues and practices			+		+					+			+				+									
Significant properties of digital objects	+	+																				+				
Preservation of e-mail			+			+																				
Preservation of text records			+			+																				
Preservation of databases			+			+		+																+	+	
Preservation of web records			+					+				+	+	+					+						+	
Preservation of spreadsheets						+																				
XML for preservation			+			+		+									+									+
Cost of digital preservation		+			+	+		+														+				
Digital repository		+			+					+		+					+	+	+	+		+		+	+	+
OAIS reference model		+		+	+			+									+								+	
Digital preservation metadata		+		+				+											+	+		+			+	