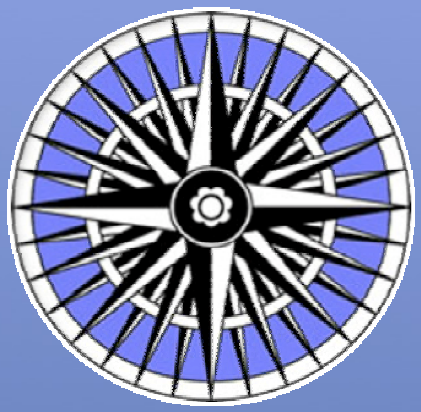
INTERNATIONAL RECORDS MANAGEMENT TRUST



Module 4

PRESERVING ELECTRONIC RECORDS

Training in Electronic Records Management

MODULE 4

PRESERVING ELECTRONIC RECORDS

Training in Electronic Records Management

General Editor, Laura Millar

MODULE 4

PRESERVING ELECTRONIC RECORDS

INTERNATIONAL RECORDS MANAGEMENT TRUST

Module 4: Preserving Electronic Records

Produced by the International Records Management Trust
4th Floor
7 Hatton Garden
London EC1N 8AD
UK

Printed in the United Kingdom.

Inquiries concerning reproduction or rights and requests for additional training materials should be addressed to

International Records Management Trust
4th Floor
7 Hatton Garden
London EC1N 8AD
UK
Tel: +44 (0) 20 7831 4101
Fax: +44 (0) 20 7831 6303
Email: info@irmt.org
Website: http://www.irmt.org

Version no. 1/2009

TERM Project Personnel

Project Director

Dr Anne Thurston, founder of the Trust, is a pioneer in defining international solutions for the management of public sector records. Both as an academic and as a programme director, she has extensive experience of working with many different governments to provide practical solutions for strengthening record-keeping systems. Her groundbreaking survey of record-keeping systems across the Commonwealth resulted in the establishment of pilot projects to restructure records systems in The Gambia and Ghana, and she established the Trust in 1989 to develop and extend this work. She joined the staff of the School of Library, Archive and Information Studies at University College London in 1980 to develop the Masters' in Records and Archives Management (International); she was also a Reader in International Records Studies. In 2000 she was awarded an OBE for services to public administration in Africa; she received a lifetime achievement award from the UK Records Management Society in 2006. She was awarded the Emmett Leahy award for Outstanding Contributions to the Information and Records Management Profession in 2007.

General Editor

Laura Millar divides her time among three careers: in archives as an archival and information management consultant and educator; in publishing as a writer, editor, and instructor; and in distance education as a curriculum developer, instructional designer, and course author. She received her MAS degree in archival studies from the University of British Columbia, Canada, in 1984 and her PhD in archival studies from the University of London in 1996. From 1994 to 1999, as Managing Editor of the Management of Public Sector Records Study Programme for the International Records Management Trust and the International Council on Archives, she was responsible for the development, testing, and delivery of 18 distance education training modules and 15 associated publications in archives, records and information management. She is the author of a number of books and articles on various topics in archives, publishing, and distance education.

Project Manager

A New Zealand born Australian based in Seattle, Washington, Michael Hoyle has a Masters degree in Information Management and Systems from Monash University in Australia. Prior to moving to Seattle in 2005, he was the Group Manager, Government Recordkeeping at Archives New Zealand. He has also worked in various information management and other roles in several government agencies in Australasia, including ten years at Archives New Zealand and six years at the National Archives of Australia. Michael has been a council member of the Archives and Records Association of New Zealand (1996 to 1999) and served the Association of Commonwealth Archivists and Records Managers (ACARM) as Deputy Chair (2000 to 2002) and as Chair (2002 to 2004). He also served the Pacific Branch of the International Council on Archives (PARBICA) as Secretary General (2002 to 2003) and President (2003 to 2004).

Module 4: Preserving Electronic Records

Authors

Adrian Brown
Shadrack Katuu
Peter Sebina
Anthea Seles

Additional Contributors

Christine Ardern
Laura Millar

Reviewers

Andrew Griffin
Michael Hoyle
Patrick Ngulumbe
Jim Suderman
Anne Thurston
Louisa Venter

The International Records Management Trust would like to acknowledge the support and assistance of the Department for International Development (UK).

Contents

Figures

PREFACE

ABOUT THE *TERM* PROJECT

This module is part of an educational initiative called *Training in Electronic Records Management* or *TERM*, developed by the International Records Management Trust as part of a wider project to investigate issues associated with establishing integrity in public sector information systems. Begun in 2006, *Fostering Trust and Transparency in Governance: Investigating and Addressing the Requirements for Building Integrity* in Public Sector Information Systems in the ICT Environment was a project designed to address the crucial importance of managing records in the information technology environment. The focus of the study was pay and personnel records, since payroll control and procurement are the two major areas of government expenditure most vulnerable to misappropriation, and payroll control is, therefore, a highly significant issue for all governments.

The project provided an opportunity to explore the management of paper records as inputs to financial and human resource management information systems, the management of electronic records as digital outputs and the links between them. It also involved examining the degree to which the controls and authorisations that operated in paper-based systems in the past have been translated into the electronic working environment.

The primary geographical focus of the study was eastern and southern Africa, and two significant regional bodies participated: the Eastern and Southern Africa Regional Branch of the International Council on Archives (ESARBICA) and the Eastern and Southern African Association of Accountants General (ESAAG). Four countries from the region (Zambia, Botswana, Lesotho and Tanzania) hosted case studies, and comparative studies were carried out in West Africa (Ghana) and Asia (India).

The products of this project, which will be available without charge, include

- route maps for moving from a paper-based to an electronic information environment
- good practice indicators to measure records management integration in ICT control systems
- these training modules on the management of records in electronic form.

The project deliverables also include case studies conducted in Botswana, Ghana, India, Sierra Leone, Tanzania and Zambia. The studies focused primarily on issues related to the management of human resources and payroll functions in governments and involved research into paper-based and computerised personnel management systems. However, they provided an opportunity also to examine records and information management in the public sector in these countries. The case studies are

most relevant to those readers focusing on personnel and payroll management. However, the findings also offer valuable insights into the challenges of automation and electronic government, and the issues involved with making the transition from paper-based to electronic records and information management. The final case studies are being made available on the Trust website at www.irmt.org.

The case studies all point to the general need for greater integration of records management in the design and implementation of electronic information and communications (ICT) systems. The good practice indicators produced by this project are intended to help governments determine whether or not records management requirements have been integrated in ICT systems and to provide a high-level guide to records management integration. The indicators are particularly relevant to Modules 2 and 3. The good practice statements that underpin the indicators are derived from generally accepted international standards but are also informed by the findings of the case studies.

It is hoped that the research conducted as part of this project will offer governments the resources they can use to increase their capacity to manage paper and electronic records as accurate and reliable evidence in electronic environments. Their ability to measure progress toward accountability will be enhanced, and there should be a higher success rate of e-governance applications.

Project Steering Team

An international steering team oversees the work of the project, consisting of the following members.

- **Stephen Sharples**, Chair of the Steering Committee, Senior Governance Adviser, Africa Policy Department, UK Department for International Development
- **Anne Thurston,** Project Director and International Director, International Records Management Trust
- **Michael Hoyle,** Project Manager, International Records Management Trust
- **Andrew Griffin,** Research Officer and UK Director, International Records Management Trust
- **Jerry Gutu,** Chief Executive Officer, East and Southern African Association of Accountants General (ESAAG) (2006)
- **Cosmas Lamosai,** Chief Executive Officer, ESAAG (2007 and 2008)
- **Kelebogile Kgabi,** Chair, Eastern and Southern African Branch, International Council on Archives (ESARBICA), and Director, Botswana National Archives and Records Services (2006)
- **Gert Van der Linde,** Lead Financial Management Specialist, Africa Division, World Bank
- **Peter Mlyansi,** Director, Tanzania Records and National Archives Department and Chair of ESARBICA (2007 and 2008)
- **Nicola Smithers,** Public Sector Specialist, Africa Region, World Bank

- **David Sawe,** Director of Management Information Systems, Government of Tanzania
- **Ranjana Mukherjee,** Senior Public Sector Specialist, Asia Region, World Bank.

More information about the project and the other deliverables can be found on the International Records Management Trust website at http://www.irmt.org/building_integrity.html.

About the Modules

The following modules have been produced as part of this project:

Module 1 *Understanding the Context of Electronic Records Management*

Module 2 *Planning and Managing an Electronic Records Management Programme*

Module 3 *Managing the Creation, Use and Disposal of Electronic Records*

Module 4 *Preserving Electronic Records*

Module 5 *Managing Personnel Records in an Electronic Environment.*

As well, the following two resources have been produced:

Additional Resources a bibliography of key resources related to the management of electronic records.

Glossary of Terms a consolidated glossary of relevant records management, electronic records management, information technology and computer terms.

These materials are primarily intended for use by records management practitioners in developing countries. The focus is on providing both a conceptual framework and practical guidance about important issues related to electronic records management. The goal is to produce a series of resources that can be used in a variety of ways, such as

- for self study
- for in-house training
- for management training institutes
- as a resource for university or college courses
- as supporting information for distance education courses.

A series of self-study questions has been included at the end of each module. These questions can be used by readers to assess their own understanding of the content provided in the module. The questions may also be used by trainers or educators to develop activities, assignments or other assessments to evaluate the success of any training offered. In order to facilitate the widest possible use of these questions by both learners and educators, they have been gathered together in one place at the end

of the module rather than interspersed throughout the text. Readers interested in developing educational or training initiatives using these modules are also directed to the MPSR training resources developed in 1999, and listed below, which offer guidance on how to adapt and use educational tools such as these.

Contributors

A number of records and information professionals were asked to contribute to the modules, including representatives from such countries as Australia, Botswana, Canada, Kenya, Singapore, South Africa, the United Kingdom and the United States. The following people have contributed to the project as contributors, editors, reviewers and production assistants.

- Keith Bastin, United Kingdom, reviewer
- Adrian Brown, United Kingdom, contributor
- Luis Carvalho, United Kingdom, administrative coordinator
- Donald Force, United States, editor
- Elaine Goh, Singapore, contributor
- Andrew Griffin, United Kingdom, contributor
- Greg Holoboff, Canada, graphic artist
- Michael Hoyle, United States, contributor
- Shadrack Katuu, South Africa, contributor
- Segomotso Keakopa, Botswana, contributor
- Lekoko Kenosi, Kenya, contributor
- Charles Kinyeki, Kenya, reviewer
- Barbara Lange, Canada, desktop publisher
- Helena Leonce, Trinidad and Tobago, reviewer
- Mphalane Makhura, South Africa, reviewer
- Walter Mansfield, United Kingdom, contributor, editor
- Peter Mazikana, Zimbabwe, contributor
- John McDonald, Canada, contributor
- Laura Millar, Canada, contributor, editor
- April Miller, United States, contributor
- Patrick Ngulumbe, South Africa, reviewer
- Greg O'Shea, Australia, contributor
- Lori Podolsky Nordland, Canada, contributor
- Peter Sebina, Botswana, contributor
- Anthea Seles, Canada, contributor
- Elizabeth Shepherd, United Kingdom, reviewer
- Kelvin Smith, United Kingdom, contributor
- Jim Suderman, Canada, contributor, reviewer
- Setareki Tale, Fiji, reviewer

- Louisa Venter, South Africa, reviewer
- Justus Wamukoya, Kenya, reviewer
- Richard Wato, Kenya, reviewer
- Geoffrey Yeo, United Kingdom, reviewer
- Zawiyah Mohammad Yusef, Malaysia, reviewer.

Relationship with the *MPSR* Training Programme

The modules are designed to build on and support the *Management of Public Sector Records* training programme, developed by the International Records Management Trust in 1999. The MPSR training resources consist of over thirty separate training tools that address basic records management issues for developing countries. While some information found in those earlier modules may also be found in this new training programme, the concept behind this new set of modules is that they build upon but do not replace those earlier fundamental records management training tools. However, this new TERM programme focuses on the electronic record-keeping environment that is becoming so prevalent in the early years of the 21st century.

Readers wishing to orient themselves to basic records management principles will want to refer back to those MPSR resources, which are available free of charge from the International Records Management Trust website at www.irmt.org. Those training resources are identified below.

Training Modules

1 The Management of Public Sector Records: Principles and Context
2 Organising and Controlling Current Records
3 Building Records Appraisal Systems
4 Managing Records in Records Centres
5 Managing Archives
6 Preserving Records
7 Emergency Planning for Records and Archives Services
8 Developing the Infrastructure for Records and Archives Services
9 Managing Resources for Records and Archives Services
10 Strategic Planning for Records and Archives Services
11 Analysing Business Systems
12 Understanding Computer Systems: An Overview for Records and Archives Staff
13 Automating Records Services
14 Managing Electronic Records
15 Managing Financial Records
16 Managing Hospital Records
17 Managing Legal Records
18 Managing Personnel Records

Procedures Manuals

19 Managing Current Records: A Procedures Manual

20 Restructuring Current Records Systems: A Procedures Manual

21 Managing Records Centres: A Procedures Manual

22 Managing Archives: A Procedures Manual

23 Planning for Emergencies: A Procedures Manual

24 Model Records and Archives Law

25 Model Scheme of Service

Educators' Resources

26 Educators' Resources

- Introduction to the Study Programme
- Glossary of Terms
- Additional Resources for Records and Archives Management
- Educators' Resource Kit
- Writing Case Studies: A Manual.

Case Studies

27 Case Studies Volume 1

28 Case Studies Volume 2

29 Case Studies Volume 3

The introduction to each module in the TERM programme includes more specific information about relevant MPSR resources that readers may wish to review in association with the TERM module in question.

A Note on Terminology

As with any material related to computer technologies, these modules contain a great deal of specialised terminology. Every attempt has been made to define key terms the first time they are used. When important concepts are discussed cross-references are included as appropriate to earlier references or to the glossary of terms. Readers are also directed to the *Additional Resources* tool for more information on various topics, and web addresses are included whenever detailed information is provided about particular organisations or specific resource materials.

The modules are written using British English (programme, organisation) though of course many computer terms use American English: thus an organisation may run a records management 'programme' but it uses a particular software 'program.' Abbreviations and acronyms are defined the first time they are used in each module and are used as sparingly as possible.

One exception is ERM for 'electronic records management': this acronym is used regularly throughout all the resources as appropriate when referring to the general concept of managing computer-generated records. When referring to an electronic

records management system – that is, to specific software programs designed to manage electronic records – the term ERMS is used. It is recognised, however, that ERMS software may also offer document management features: supporting the creation, use and maintenance of both documents (such as works in progress) and records (official, final documents). When referring specifically to software that manages both documents and records, the acronym EDRMS is used, but the acronym ERMS is used more often, particularly when the concept of electronic records management systems is discussed more generally.

For More Information

For more information or to download a copy of these resource materials free of charge, go to the International Records Management Trust website at www.irmt.org. The Trust can be reached as follows:

International Records Management Trust
4th Floor
7 Hatton Garden
London EC1N 8AD UK

phone +44 (0) 20 7831 4101
fax +44 (0) 20 7831 6303
email info@irmt.org
website www.irmt.org

INTRODUCTION

This module is the most technically challenging of the five modules in the series. Not only are the issues involved complex, but the solutions to preserving electronic records over long periods are still emerging. Please note that this module is not a 'how-to' manual but rather an educational guide, designed to introduce readers to key concepts and ideas and direct them to further sources of information.

This module examines the nature of, and challenges associated with, the preservation of electronic records. The purpose of archival preservation is to ensure that records remain accessible over time in such a way that they can be considered authentic and reliable evidence. Not only must records be accessible, but their intrinsic value must also be retained. For traditional manual records and paper-based collections, including textual and audiovisual records created before the advent of computer technologies, the principal obstacle to preservation is the physical decay of the materials themselves. Paper records can become damaged through excessive handling and as a result of deterioration caused by the acids in the paper fibres, leaving documents brittle and discoloured over time. As well, the colour dyes in photographic films and prints continue to be chemically active and can fade when exposed to excessive light or high temperatures.

The task of preserving electronic records over long periods presents a number of complex challenges. As discussed in Module 1, digital information is stored in the form of bits: ones and zeros that denote values in binary notation. These bits have no inherent meaning; rather, they represent the encoding of information according to a predefined scheme. Computerised information can only be read with the help of special computer hardware and software capable of translating that information into human-readable form.

A significant challenge for preserving electronic records is the degradation of the software or systems required to make digital information readable. Another difficult problem in preserving electronic records is the inevitable obsolescence of the technology used to create them. For example, some digital photographs are created or saved in a popular computer format called TIFF or 'tagged image file format.' In order to view that image, the user needs access to TIFF image viewing software to render the bits into viewable form. That image could then be converted into another file format, such as GIF or 'graphic interchange format.' The image the viewer sees will look identical to the TIFF image, but the computer is reading two completely different records, each with its own unique qualities.

Whichever format is chosen, the user needs access to multiple computer technologies: the appropriate image rendering software; the right operating system and hardware configuration to view the image; and the hardware and software required to run the

computer in the first place. As well, the user will need a way to connect the computer to the media on which the image is stored, such as a computer's hard disk drive or a CD-ROM storage device, which might require specialised software to function.

As a result, access to any digital object – photograph, document, database, spreadsheet or other electronic information resource – depends upon a complex network of interconnected technologies. This network is called a 'representation network' since it must comprise all of the elements required to represent the object correctly. The absence or failure of any part of this network could render the object inaccessible.

One of the difficulties in securing a suitable representation network is that computer technologies are constantly changing and evolving. Information technology is a rapidly advancing field, and new and improved technologies are being developed regularly. Equally, economic pressures force technology developers to follow a regular cycle of product replacement. New products and new versions of existing products are brought to market regularly, and as new products become available, existing products cease to be supported.

The currency of any given computer technology is, therefore, typically very short: perhaps five to ten years at most. This rapid rate of obsolescence applies to all technologies in the representation network, including file formats, software, operating systems and hardware. The challenge of digital preservation, therefore, lies in maintaining a way to access digital objects in the face of rapid technological obsolescence. In particular, digital preservation requires methods for identifying and predicting the impact of technological change on digital collections and for executing appropriate preservation strategies to reduce this impact, often even before the records themselves have been created.

Maintaining accessibility in itself is not enough – in an archival context, the authenticity of the record must also be preserved. As discussed in Module 1, the authenticity of an electronic record derives from three essential characteristics: reliability, integrity and usability. But authenticity in a digital environment is complicated by the fact that the preservation of electronic records always entails some form of transformation. Digital preservation requires the management of objects over time, and the techniques used may result in frequent and profound changes to the technical representation of that record. Over time, new technical manifestations of a record will be created, making it that much more important to confirm the authenticity of the record.

Therefore, ensuring the preservation of and access to electronic records involves understanding some of the important technological and management issues associated with digital preservation.

Digital preservation represents a formidable challenge, at both a technical and organisational level, and many challenges remain to be fully resolved. However, these difficulties should not be seen as obstacles to the establishment of practical preservation policies. Every element of the preservation process described in this

module can be addressed with varying degrees of complexity as suits the resources available.

Furthermore, the widespread and active research activities in this field should be grounds for optimism – more mature and integrated solutions should become widely available in the next few years. But the rapid pace of research should not be seen as a reason to postpone action now. Even electronic records created relatively recently can be under threat, and the adoption of some simple preservation and security measures can significantly reduce these risks immediately.

This module provides an overview of the issues involved with preserving electronic records over time and examines some of the options for establishing a preservation strategy. The discussion focuses on preserving digital objects since, as discussed in Module 1, the preservation of a complex electronic record may well involve ensuring the protection of many different components or objects.

Unit 4.1 examines important principles associated with the task of digital preservation, including considering the characteristics of digital objects; understanding the role of different characterisation software programs, and addressing different types of preservation, including refreshing data, replicating data, migration and emulation. The unit includes a brief overview of issues consider when choosing the best preservation strategy.

Unit 4.2 introduces different preservation practices, including developing a preservation policy, preparing a risk assessment, establishing security and access controls, ensuring the integrity of the electronic record, managing metadata, managing the content of electronic records and planning for emergencies.

Unit 4.3 examines the role, purpose and nature of a digital repository. Specific topics covered include the concept of a trusted digital repository, requirements for establishing and maintaining a trusted digital repository, selecting hardware and software solutions, understanding the ingest process and the concept of information packages, choosing storage devices, preparing records for preservation in the repository, ingesting records into the repository, deciding when to destroy original records, monitoring the status of the preservation programme and staying current with technological changes.

Unit 4.4 introduces information about various research projects currently underway around the world on the subject of electronic records management and offers comments on possible future directions and trends in electronic records preservation.

At the end of the module is a series of study questions that readers may wish to review in order to help them reflect on the topics discussed throughout the text.

FOR ADDITIONAL INFORMATION

This module more than all the others in this training programme requires that readers turn to other resources for more detailed explanations of issues and actions. Therefore, readers should pay particular attention to the relevant citations in the *Additional*

Resources document to find references to publications, websites, associations and other resources relevant to the general topic of electronic records preservation.

Readers are also reminded that the *Glossary of Terms* includes definitions for key records management terminology. Readers wishing to study some of the fundamentals of records management as related to this specific topic may wish to review some of the MPSR training modules, available online at www.irmt.org.

Of particular relevance to the preservation of electronic records are the following MPSR products:

Training Modules

- Preserving Records
- Emergency Planning for Records and Archives Services
- Managing Electronic Records

Procedures Manuals

- Planning for Emergencies: A Procedures Manual

Case Studies

- Pitt Kuan Wah, Singapore, Preserving Electronic Records at the National Archives of Singapore: a balancing archival act and a shared responsibility
- Roger Craig, Cayman Island, A Disaster Preparedness Plan for the Cayman Islands National Archives
- Cassandra Findlay, Australia, Development and Implementation of the Immigration Department's New International Traveller Movements System
- Pino Akotia, Ghana, Management of Financial Records: The Ghana Case Study
- Ann Pederson, Australia, Storage/Preservation Case Study: Responding Effectively to a Disaster.

Readers are also directed to the work of various national, state and regional archival institutions that are actively developing strategies for the management and preservation of electronic records; specific information about some of those initiatives is provided throughout this module.

Laura Millar, the editor of this module, acknowledges with particular thanks to information provided by the staff of the Washington State Digital Archives in Spokane, Washington, who provided an extensive tour of their facility in Spokane, Washington and also offered detailed information about the operations of the digital archives during a site visit in November 2008. Their support for this project is gratefully acknowledged.

UNIT 4.1

UNDERSTANDING KEY CONCEPTS IN DIGITAL PRESERVATION

Module 1 introduced some important concepts related to electronic records, including the opportunities and challenges brought by electronic information technologies; the nature of computers and electronic records; the technological, legal and organisational frameworks for electronic records management; the importance of record-keeping standards in the preservation of electronic records as authentic evidence; and the importance of metadata as a mechanism for preserving the authenticity and integrity of electronic records.

The information in this unit expands on those concepts, identifying some principles and ideas specifically associated with the task of digital preservation. The term 'digital preservation' is used to refer to the overall approach to preserving information and records created using computers, including electronic records. The issues raised include: the characteristics of electronic records and the fact that they are composed of different digital objects; the role of different software programs for identifying the characteristics of those different digital objects that make up electronic records; the difference between active and passive preservation of electronic records; and different types of preservation, including refreshing data, replicating data, migration and emulation. The unit ends with a brief overview of issues consider when choosing the best preservation strategy.

The Characteristics of Electronic Records

As discussed in Module 1, electronic records may consist of many different components or elements. To repeat the definition provided in Module 1, an electronic record element is

> any component of information created electronically that forms part of an electronic record and that is usually stored separately within the digital file making up the electronic record as a whole.

Every electronic record consists of at least one digital object, component or element, such as the bits of data that come together to create a word processed document. And some electronic records, such as photographs, video clips or web pages, may contain many different objects or elements.

One of the key steps in preserving any electronic record is to identify the precise characteristics of the record, including all the objects or components that make up that record. The term for this process is 'characterisation.' If the precise technical qualities

of a digital object are not sufficiently understood, it is impossible to preserve the record in an accessible, authentic form. It is necessary, therefore, to understand the significant technical properties of any digital object so that these properties can be preserved over time.

The action of 'characterisation' involves identifying the record, validating the record and extracting core metadata about the record, as generated at the time the record was originally created and actively used.

Identifying the Record Objects

Identifying the record involves identifying and documenting the precise computer file format and version of the digital object(s) to be preserved. For example, in order to identify a word processed record precisely, it is necessary to identify the exact name and precise version of the word processing software used to create the record, along with information about the record, such as its title. The process of identification answers the question 'in what format is this particular digital object created and stored?' Is it a word processed document? What software was used? Is it a digital photograph? What type of camera was used, when was the picture taken, how was it transferred to digital storage in a computer?

Understanding and identifying file formats, as discussed in Module 1, is important to identifying the record. For instance, the characteristics of one kind of computerised record, such as a TIFF file holding a digital photograph, will be different from another type of computerised record, such as a JPEG file holding the same digital image. An MP2 audio file and an MP3 audio file will have different characteristics. Since there are hundreds of file formats, it is useful to identify those formats that are commonly used within the organisation, as these will be the ones requiring the most attention during the preservation process.

Validating the Record Objects

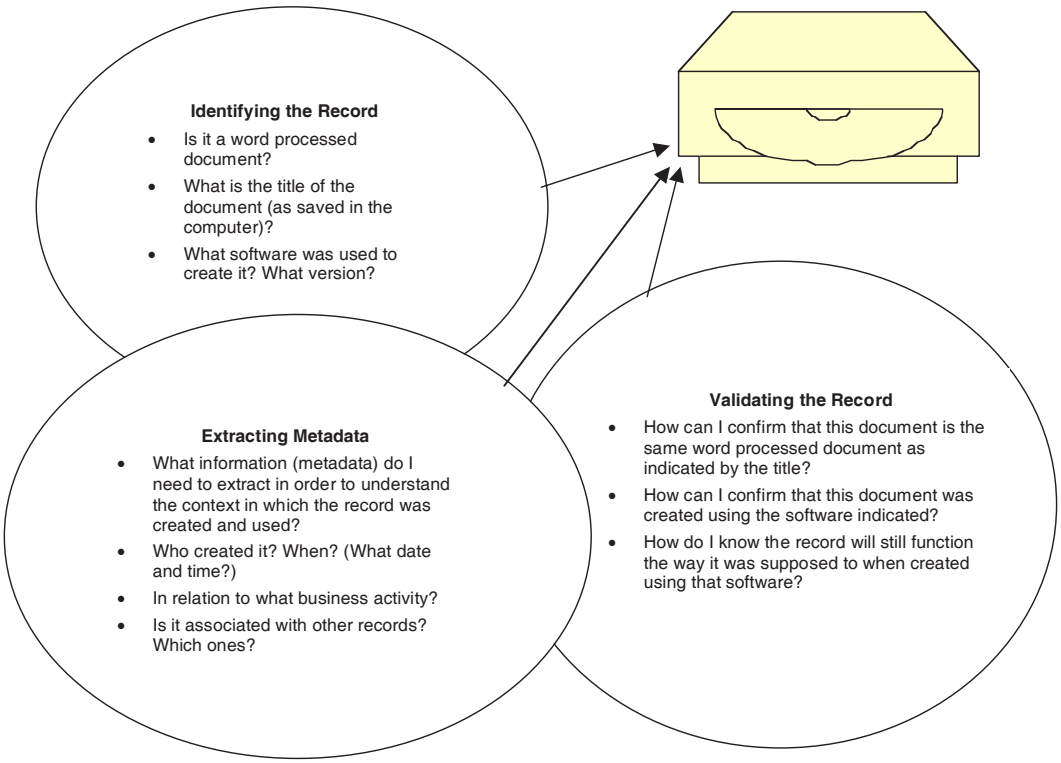
It is also important to confirm that the digital object(s) still retain the properties given when first created. For instance, to preserve a word processed document created in Microsoft Word 2000, it is necessary to know that the record still exists as a MS Word 2000 document and has not been saved as a MS Word 2007 document or a text file or an ASCII file. 'Validating' the record object should answers questions such as: 'if this object is supposed to be in format A, can I confirm that it is in that format?' and 'if this object is in format A, will it still perform the functions originally designed for format A?' These functions might include, for example, sorting of text, performing mathematical calculations or showing specific highlights or text formatting.

Extracting Record Object Metadata

It is important to extract certain information from a digital object and to store that information as metadata, to help understand how to manage that record. For example, knowing the type of compression (reduction in data volume) used to store a digital image can be essential for determining an appropriate method of migration. Similarly,

the date on which a digital photograph was captured, which is often automatically recorded by the camera within the image file, is an important part of the record. Figure 1 below illustrates the idea that electronic records are made up of different elements or components.

Figure 1: Characterising the Objects that Make Up an Electronic Record



Characterisation Software

Characterising the objects or components of an electronic record sounds like a daunting task. Fortunately, a number of freely available software tools (called characterisation software) have been developed that can be used by organisations to support the task of characterising digital objects. Tools such as the few software products identified below can be used to generate technical metadata which will assist in the preservation of the digital objects that make up electronic records.

DROID and PRONOM

DROID stands for Digital Record Object Identification. DROID is a software tool developed by The National Archives of the UK; its function is to perform automated identification of file formats for large groups, or batches, of digital objects. DROID is available free for use by any organisation wishing to preserve digital objects, to help the organisation identify the precise format of those objects.

To facilitate the management of these digital objects, The National Archives has also developed an online information system called PRONOM, which maintains

information about different data file formats and related software. PRONOM was first developed for use within The National Archives, but it is now available as a free resource for anyone needing information about different software products and associated file formats.

So, for instance, once an organisation has used a tool such as DROID to identify the file formats for different digital objects, the organisation can use PRONOM to find out the current status of the software products that had been used to create the different records that contained those digital objects, and PRONOM will identify any technical requirements associated with preserving the records created using those software products. In that way, the organisation can identify a record and understand what actions need to be taken to ensure the record is preserved so that it remains usable in the future.

More information about DROID can be found at the official website at http://droid.sourceforge.net/wiki/index.php/Introduction and more about PRONOM can be found at the official website at http://www.nationalarchives.gov.uk/pronom.

Jacksum

Jacksum, made from the words JAva and CheCKSUM, produces numerical checksums and compares these with numbers calculated at different dates. (Java is a type of computer programming language, and a checksum is a mathematical process for detecting errors in computer data. Checksums are explained later in this module.) This software program is available free.

More information about Jacksum can be found at the official website at http://www.jonelo.de/java/jacksum/index.html.

JHOVE

JHOVE, which was introduced in Module 3, is a software program that supports the identification, validation, and characterisation of digital objects. JHOVE (pronounced 'jove') will automatically identify the format of computer files; verify that the format extension and file are the same; and extract technical metadata from the file.

More information about JHOVE can be found at the official website at http://hul.harvard.edu/jhove/.

NLNZ Metadata Extractor

The National Library of New Zealand has developed a tool for extracting metadata from a number of formats. The NLNZ Metadata Extractor extracts metadata from different file formats, including PDF documents, image files, sound files, Microsoft Office documents and others. The metadata gathered may relate to the hardware or software used to create the file, the date and time it was created, or the name or title of the person responsible for creating or using the file. This information can then output that information into XML or other open source formats so the information can be used to determine preservation criteria and actions.

> More information about NLNZ Metadata Extractor can be found at the official website at http://meta-extractor.sourceforge.net/.

Planets

The Planets project (Preservation and Long-term Access through Networked Services) is a four-year project co-funded by the European Union to address digital preservation issues. The goal is to develop tools and services to help ensure long-term access to digital resources. The project aims to develop a series of technical registries, together with new preservation planning tools designed to assist organisations to identify preservation options.

> More information about Planets can be found at the official website at http://www.planets-project.eu/.

Persistent Identifiers

An important concept in digital preservation is the idea of a 'persistent identifier.' A persistent identifier is a unique name or code that is assigned to a digital object; that identification code can then be used in perpetuity to refer to and retrieve that particular object. Many different organisations within the information management community are working on the development of standards for persistent identifiers, particularly for the management of web-based information. Among the various initiatives underway to develop criteria for assigning persistent identifiers are the following:

- the Uniform Resource Name (URN)
- the persistent URL (PURL)
- the Handle system
- the digital object identifier (DOI)
- National Bibliography Numbers (NBNs)

- the Archival Resource Key (ARK)
- the Open URL.

Any organisation wishing to establish sound digital preservation practices needs to find a way to create persistent identifiers for its electronic records and objects, most likely by adopting one of the standards under development. That unique code will then become attached to the record's descriptive metadata, becoming the tool for finding and retrieving the record over time.

Passive and Active Preservation

There are two overarching approaches to the preservation of digital materials: passive preservation and active preservation.

> **Passive preservation** is the process of ensuring continuing integrity of, and controlled access to, digital objects along with their associated metadata. Essentially, passive preservation aims to 'keep' the original digital object intact without changing the technologies used to store or process it.
>
> **Active preservation** seeks to ensure the continued accessibility of electronic records over time by actively intervening in how records are stored and managed. Active preservation involves 'moving' the digital object into a new storage environment, which may depend on new technologies that were not in existence when the object was originally created and used.

Both passive and active preservation require that the integrity of the original digital object be protected. This integrity is protected either by preserving the original digital object just as it was when it was created, or by recreating the essence of the object using new and different technologies from those originally used. As stated by the UK National Archives in its 2006 report, *Generic Requirements for Sustaining Electronic Information Over Time,* authenticity does not mean preserving the actual original object as it existed when first created. Instead, the National Archives argued, 'a record is considered to be essentially complete and uncorrupted if the message that it is meant to communicate in order to achieve its purpose is unaltered.'

Passive preservation is often carried out by one of three actions: refreshing data, replication or emulation. Active preservation is often carried out by the process of migration. Each of these approaches is described below. As will be seen, not all of these approaches work for the preservation of all types of records, and some may be more effective when used in conjunction with others. However, it should be recognised that refreshing, replication and emulation are usually used as short-term measures for preserving electronic records during their active use in an organisation, whereas migration is the more common approach to preserving records as part of a formal digital preservation programme. Whenever possible, organisations should consider developing active preservation programmes as a priority if the goal is long-term preservation.

Refreshing Data

Refreshing is the process of copying data from one medium to another of the same type. During the process of refreshment, the hope is that the bits of data do not change. The purpose of refreshment is to replace data in one medium with a copy that is sufficiently the same that the data can continue to be accessed without difficulty. For example, refreshment may consist of copying membership lists from an old floppy disk onto a CD-ROM disk so that the data can be accessed using the same database management software.

Refreshing is necessary because storage media deteriorate and because the hardware needed to access and use data may change, meaning the storage media can no longer be used. For instance, people used to store documents on 3 ½" floppy disks, but as computers stopped being manufactured with those disk drives, people had to copy the data onto other media, such as CDs or hard drives.

The periodic need to refresh electronic records onto new media is inevitable given the continuous changes in computer storage media. However, selecting the best media available can reduce the frequency for refreshing data, since high-quality and stable storage media should remain usable for a longer period. Records professionals are advised not to 'jump on the bandwagon' of new media technologies too quickly, since the media chosen may not in fact be sustainable in the long term.

For example, 'zip drives' became popular in the mid-1990s because the disks provided greater storage capacity than the floppy disks commonly used at the time. By 2004, however, sales had decreased considerably, as more and more people used CD-ROM and DVD technologies. Computer manufacturers stopped building computers with zip drives, and it became harder to access the storage media. Today, CDs and DVDs are popular and common storage devices but it is always possible a new technology will replace them, in which case it will be necessary to refresh the data by moving everything to a newer storage medium.

In order to assess the status of different storage media, records professionals should remain aware of changes and developments in computer technologies. In 2003, the National Archives in the UK published a guidance note for selecting storage media, called *Digital Preservation Guidance Note 2: Selecting Storage Media for Long-term Preservation.* Other archival agencies are regularly updating information about which media are more or less suitable for the storage of electronic records with enduring value.

A copy of the guidelines can be downloaded from the TNA official website at http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf. See *Additional Resources* for more information on selecting storage media and refreshing data.

While refreshing is not supposed to involve changing the software used to read the data, it is often necessary not just to refresh data but also to migrate data to new systems, in order to make it accessible with new computer programs. Migration is discussed below.

Whenever data is refreshed, the files that have been moved should be verified at the bit level by completing a checksum or other verification or validation process, as described later in this module. This validation process is intended to ensure that the content has been copied without corruption or loss.

Replicating Data

Replication is a similar process to refreshment, but with one difference: the location where the record is stored will likely be different when a file is replicated. Again, the goal of replication is to ensure the bits of data do not change. Data that exist in only one location are highly vulnerable to damage or loss. The software or hardware could fail; someone could alter or delete the files accidentally or intentionally; or the data could be lost in a fire, flood or other environmental disaster. Replication helps ensure the survival of information, by storing the files in several different locations.

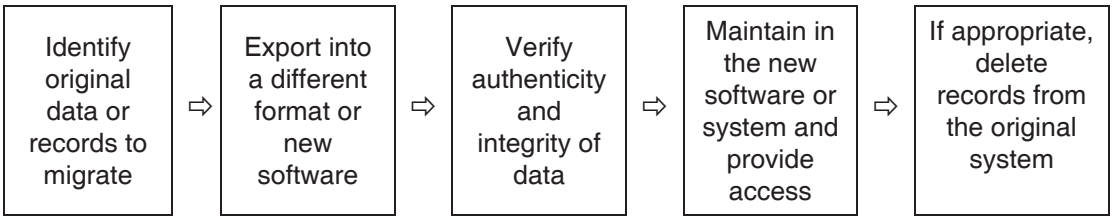
Replication is different from refreshing data, since the new copy of the electronic record is stored in a different place. Replication is also different from the process of backing up data, since replication may involve copying specific electronic records, whereas backup processes usually involve copying entire systems, with software and data together. Accessing replicated files requires knowing what software and hardware were used to create the records in the first place, which makes the preservation of metadata so important.

The existence of replicated electronic records can cause more difficulties than it can solve. If too many copies of data exist, it is much more difficult to monitor processes such as version control, migration and access. Detailed information needs to be kept about what has been replicated and where it is stored, and the organisation needs to decide how often it will replace copies with more up-to-date copies, so that it does not end up with multiple copies of information, some of it superseded, stored in various locations inside and outside of the office environment.

Migration

One method of active preservation is known as ‘migration.’ Migration is the process of translating data or digital objects from one computer format to another format in order to ensure users can access the data or digital objects using new or changed computing technologies. As discussed later, migration is the common method use to transfer records into digital storage repositories as part of a formal preservation program. During migration, the bits of data may change, which can risk the integrity of the data. Figure 2 below illustrates the key steps involved with migration.

Figure 2: Key Steps in Data Migration



Migration-based preservation strategies are similar to refreshment, in that both approaches involve converting the digital object, rather than the technology used to create it, to a form that can be accessed in a contemporary environment. The concept of converting an object from one format to another is widely understood by anyone who uses, for example, word processing software such as Microsoft Word. The 'Save as…' option in software tools such as MS Word provides most users with their primary experience of migration: the option allows users to save a digital object – such as a word processed document – in a format other than the one in which it was created.

The MS Word example is a very basic approach to migration. It is important to remember that while the principle of migration may seem straightforward, the practice can provide significant challenges. The formats in which digital objects are created and used vary enormously, and even the formats used for specific types of object, such as word-processed documents, can vary significantly in their functionality. In part, this diversity is a result of advances in technology, but it also reflects the efforts of software developers to establish the uniqueness of their particular products and to retain their market share by regularly releasing new versions of products with different features and improvements.

As a result, there is rarely a precise match between the features of the source and target formats in any migration process. The loss of information and functionality is a very real possibility. Furthermore, the conversion process itself may not be implemented effectively, and further loss may occur.

A number of different approaches to migration have been advocated, including normalisation, migration at obsolescence and migration on demand. The primary difference among the different approaches is the timing of their use. Understanding each of these approaches to migration is essential to knowing how to approach digital preservation within an organisation.

Migration by Normalisation

Normalisation is sometimes referred to as 'migration on ingest.' (The process of transferring records to a digital storage repository is referred to as 'ingest.') Normalisation involves migrating a digital object from the original software into an open source, standards-based format so that it can be used without having to rely on the original, possibly proprietary, software system used to create it. Open source refers to software for which the source code is freely available. Normalising seeks to

minimise the frequency and complexity of future migration cycles by going straight to an open source format that, ideally, will always be available and accessible.

It must be remembered that a normalised record is not 'the original' record. Some information may be lost during the process of normalisation. However, there is a general belief that normalisation allows digital objects to be preserved longer, because they are no longer held in commercial software systems and are stored in formats bound by accepted standards. The goal is to ensure that the normalised digital object 'performs' as much as possible like the original while not being dependent on the technology originally required, which may be too expensive to preserve over the long term.

Normalisation is an important feature of any cost-effective digital preservation programme, because it requires fewer resources – financial and otherwise – to ensure that electronic records are maintained and preserved. This is because, for example, normalisation avoids issues related to copyright and the use of proprietary software since normalisation as a rule opts for open-source solutions. However, no one normalisation format will meet all the requirements of an organisation. Therefore a combination of approaches may be necessary.

Normalisation Software

Research has been underway for some time to create software solutions that will facilitate the process of normalising records. Three products are described here: XENA, PDF/A and METS.

XENA

XENA stands for 'Xml Electronic Normalizing for Archives.' XENA was developed by the National Archives of Australia to help in the long-term preservation of electronic records created by the Australian Government. XENA converts some formats into .xml so that the formats can then be viewed by the XENA viewer.

To use XENA, organisations will benefit from installing OpenOffice, which is a suite of open source software programs for word processing, spreadsheets, presentations, graphics, databases and other applications. OpenOffice works on all common computers and is available in several languages. Regardless of whether the organisation uses OpenOffice, however, by converting files into a binary code that is not tied to a software program, organisations can then reconstitute the document at a later time using other software systems. It should be noted that, at this time, sound files, databases and dynamic, interactive or experiential files cannot be normalised by XENA.

For more information on XENA, see the official website at http://xena.sourceforge.net/.

PDF/A

Another normalisation format is called PDF/A, which is based on the PDF format that was developed by Adobe Systems in 1993. Similar to read-only PDF files, PDF/A files retain the visual appearance of the original record but do not allow the record to be changed. Like PDF files, PDF/A records work for visual records only; audio and video files and complex formats such as moving images cannot be supported.

To create PDF/A files, the organisation needs to acquire software tools similar to those used to create PDF files. Adobe Acrobat and Microsoft have products available, but it is important to ensure that any product used conforms with the ISO Standard 19005-1: *Document Management: Electronic Document File Format for Long-term Preservation,* which dictates the format in which PDF/A files must be preserved to be considered authentic and reliable records.

The ISO standard also includes requirements for the software products used to read PDF/A files, including ensuring that colours, fonts and annotations can be accessed without any degradation of the original formatting. Therefore, a PDF/A document cannot contain information or formatting that relies on access to external sources, such as hyperlinks to web pages and so on.

For more on PDF/A, see the official website at http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

METS

Finally, for more complex records such as databases, which cannot be normalised by XENA or PDF/A, it is possible to normalise records to what is called a METS file. METS is an acronym for Metadata Encoding and Transmission Standard; the METS system separates complex records into their constituent parts and 'encapsulates' them with the necessary .xml metadata to allow the object to be reproduced later in its 'original' form.

A METS document consists of the following components:

- a METS header, describing the METS document itself
- descriptive metadata about the record, including information about the content, context and structure
- administrative metadata about the record, including information about how the files were created and stored, intellectual property rights, provenance and so on
- a list of all files that, together, comprise the digital object
- a map that outlines the hierarchical structure for the object and links the different elements to the content files and metadata associated with each element

- structural links between the different components outlined in the structural map
- a behaviour section, that can be used to link behaviours or executions with the relevant element in the digital object, so that if the object needs to carry out a particular 'behaviour' it has the instructions to do so.

> For more information on METS, see the official website at http://www.loc.gov/standards/mets/.

Migration at Obsolescence

Migration at obsolescence lies at the opposite end of the spectrum to normalisation. Sometimes known as 'just in time' migration, this approach advocates that objects be migrated only as and when dictated by technological obsolescence: that is, when they are about to become inaccessible. Records can be migrated to new file formats or to current versions of old formats or they can be migrated to open-source formats through normalisation. Waiting for records to age before migrating them is dangerous, however; the records may have become damaged during a long period in computer storage systems, reducing the quality and integrity of the 'original' at the time of migration.

Migration on Demand

The strategy of migration on demand lies between these two extremes of normalisation and migration at obsolescence. Migration on demand involves storing digital objects in their original formats and only migrating them to current formats 'on demand,' such as when a user needs access to particular objects. Migration on demand can be seen as an ad hoc approach, since no planning has gone into deciding what should be migrated or not. Perhaps a user wishes to see a number of records for a specific reference purpose, but there may be many other records of great legal or administrative significance that are not needed at that time. The records requested by the user will get priority, when in fact the other, equally or more important records could be at greater risk of loss or damage.

Emulation

Emulation is the process of using one computer device or software program to imitate the behaviours of another device or program, thereby obtaining the same results when accessing or using digital objects. Emulation strategies use software or hardware – called the emulator – to recreate the functionality of obsolete technical environments on modern computer platforms. During emulation the bits of data are replicated and are not exactly as they originally were; the loss of information is a distinct possibility. However, emulation does allow access to the original object as though it were still housed in its original computer environment. For example, special software can be used on a present-day personal computer to produce exactly the appearance and

behaviour of a document, such as a presentation, that was created on an older computer using software that is no longer in use. In other words, one piece of technology is allowing the computer to act as though it were another piece of technology.

Proponents of emulation strategies argue that emulation delivers the most authentic possible rendition of a digital object. Critics of emulation, on the other hand, express concern about the significant technical challenges involved in developing emulation technology as well as the difficulty of establishing whether or not, in the end, the user is left with completely authentic recreation of the original object. In practice, it can be difficult to emulate the exact behaviour of an old system, especially when it is not fully documented. Critics also warn about potential limitations faced by future researchers who need to use obsolete technologies to access records. A number of different approaches to emulation are possible, as described below.

Software Emulation

By emulating a software package, any digital object viewable by that software can be accessed. For instance, by emulating Microsoft Windows 2001, it would be possible to open documents created several years ago using that software. Software emulation requires the development of separate emulators for every desired software package, which can be time consuming, costly and complicated.

Operating System Emulation

Rather than develop emulators for each type of software used, it is possible to develop an emulator for a specific operating system (for example, one that is no longer in use) so that all the original software that was used by that operating system can still be accessed. Emulating an entire operating system offers significant advantages because it is not necessary to develop a great number of separate emulators. However, emulating an entire operating system can be very complicated and costly. (One example is the software Virtual PC, which users of Apple computers can use to emulate a 'traditional' personal computer environment. For more information about Virtual PC, see http://www.microsoft.com/windows/products/winfamily/virtualpc/default.mspx.)

Hardware Emulation

Operating systems require specific hardware platforms, the hardware components that make up the computer itself. The next logical step, therefore, is to emulate the hardware underlying any operating system or software program. For example, contemporary PC hardware is likely to be able to support a wide variety of operating systems, including various versions of MS-DOS, Windows, UNIX and Linux. By emulating a specific hardware platform, it is possible to install and run any original operating system supported by that hardware. This level of emulation further reduces the number of different emulators required, but the increased level in technical difficulty may be a much greater challenge.

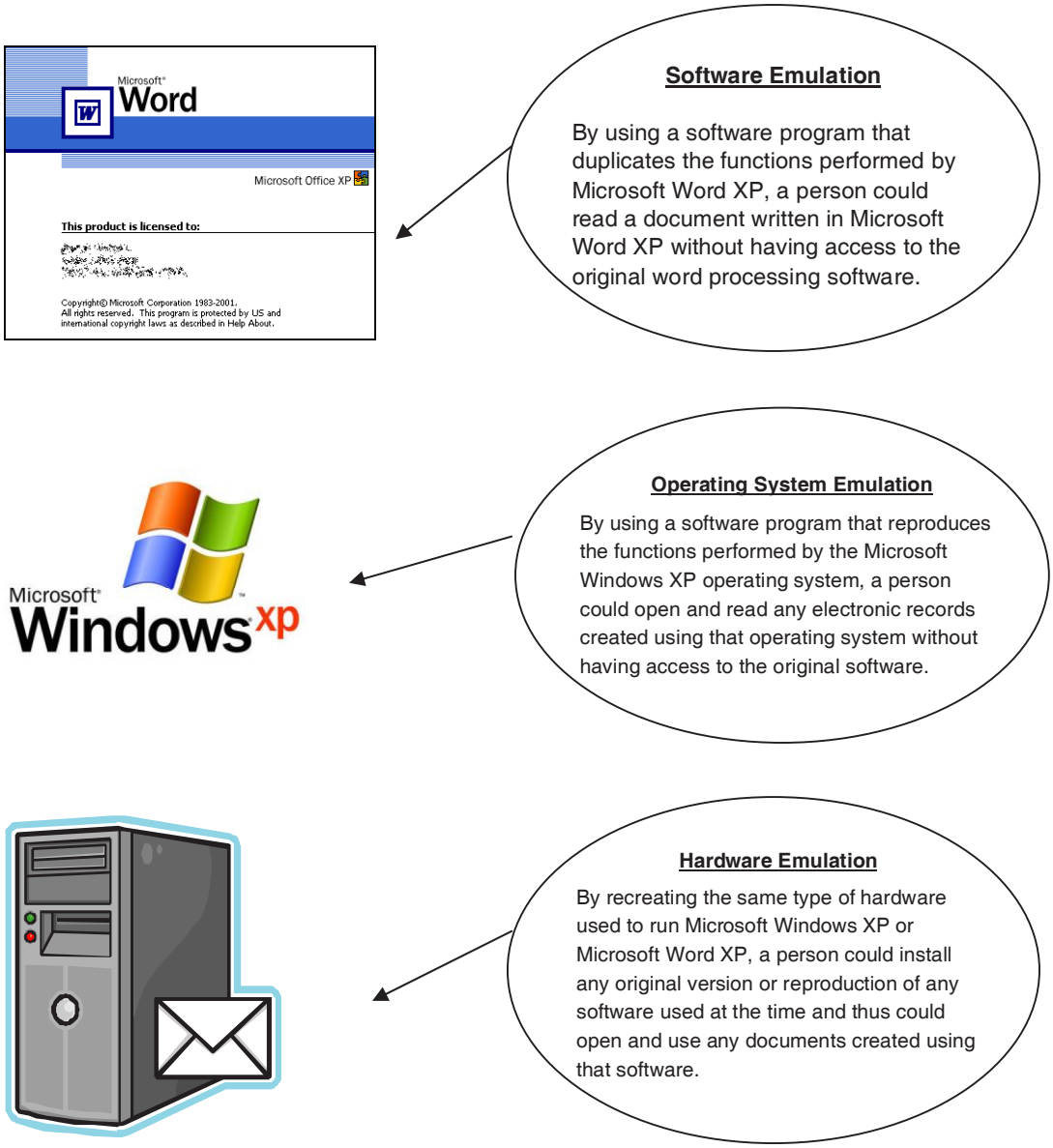
Drawbacks to Emulation

The use of emulation creates a major technical dependency upon the emulator itself. To maintain access to the emulated objects, either the emulator itself must be preserved or new emulators must be periodically created for the latest computer platforms, operating systems and software. Emulation also requires preparing solutions before problems have been diagnosed, meaning that there may be high initial costs in terms of research and experimentation. For example, in the 1980s the British government developed an emulation process while capturing digital images of pages from the *Domesday Book,* the famous record of the great survey or census conducted in England in 1086. Unfortunately, over time it became more and more difficult to maintain the software and hardware needed to access those images. In 2006, the British government developed new technologies to ensure it could continue to access this historic document.

For more information on the Domesday Book project, see The National Archives' web page at http://www.nationalarchives.gov.uk/preservation/research/rescue.htm.)

Figure 3 below shows some of the environments that might be considered when deciding what records to emulate. Note that while Microsoft examples have been shown, that is only one commercial product that may be involved in emulation; both commercial and non-commercial software and operating systems may be candidates for emulation.

Figure 3: Computer Environments That May Be Emulated



Choosing the Best Preservation Strategy

Refreshing data and replicating data are usually considered only temporary strategies at best. They offer short-term benefits but are not considered useful for long-term preservation of digital objects. Refreshing data is important if there is a risk that the storage media is deteriorating or at risk of becoming obsolete. Replicating data is valuable to ensure that multiple copies are available in case of an emergency, as long as all copies are clearly identified and tracked.

Migration and emulation, on the other hand, are both strategies that have long-term value. They are each useful for different purposes. There may be reasons why one is

more suitable in one instance, and another in another instance. Still, the software used to capture and preserve electronic records in digital storage repositories requires that the records be transferred from the original software environment into an open-source or non-proprietary software environment.

Both migration and emulation are reasonable options for action, and both may be used by the same organisation, depending on the types of records it creates, the computer technologies it uses and the demands placed on it for access to and preservation of digital information.

Now that some core concepts of digital preservation have been outlined, the next task is to consider what steps can be taken immediately in order to improve the physical condition in which electronic records are stored and managed. Basic digital preservation practices are outlined in the next unit.

UNIT 4.2

BASIC DIGITAL PRESERVATION PRACTICES

Preservation is an ongoing process. There is no end point to digital preservation, unless a digital object ceases to be considered worthy of preservation. This fact is equally true in the world of traditional preservation, although it may be less apparent because of the much greater timescales between preservation interventions in the manual environment. For example, a paper record may be safely stored for 100 years or more in an acid-free file folder in a stable storage environment. The work involved in renewing preservation strategies for that paper record might only take place once every century, if the record is moved to a new storage facility or placed in a new acid-free folder.

Preservation of electronic records requires much more intervention, and it involves the expertise of both records professionals and technology specialists. If preservation actions do not begin early, it might not be possible to preserve the electronic record, or restore it and use it, five years from now, never mind a century from now.

Ideally, an organisation will establish a formal digital preservation programme for all valuable electronic records. Understanding preservation strategies and implementing a preservation programme are discussed in more detail in the next units. But first, it is worth considering basic steps that can be taken in order to establish a measure of control over digital objects in the short term.

Protecting the authenticity of the electronic record involves performing several activities, outlined below, including establishing security and access controls; ensuring the integrity of the record; managing metadata; managing storage media; managing the content of digital objects; and planning for emergencies. These activities are important regardless of whether a formal preservation programme has been established. In other words, even if the organisation has not decided to establish specific preservation procedures (such as refreshing, replication, emulation or migration) as discussed later, it can take steps to improve the chance that digital objects will be protected, at least until more comprehensive approaches can be instituted.

This unit briefly examines a range of fundamental preservation practices, including: developing a preservation policy, preparing a risk assessment, establishing security and access controls, ensuring the integrity of the electronic record, managing metadata, managing the content of electronic records and planning for emergencies. Many of these issues are addressed in much more detail in other useful publications; see the Additional Resources tool for references.

Readers are directed to the *Additional Resources* document for more information about electronic records preservation.

Developing a Preservation Policy

A clearly documented and realistic preservation policy is an essential foundation for any sustainable digital preservation programme. The information provided in this module will highlight the range of issues that need to be addressed in a policy. Particular attention should be paid to the following issues, which should be articulated as specific guidance in the policy itself:

- the benefits of digital preservation
- the scope and objectives of a digital preservation programme
- the legal, technical, business and infrastructure requirements for a successful digital preservation programme
- the costs and resources required to maintain the programme
- the different roles and responsibilities for digital preservation
- the scope of the programme (such as national, institutional, international, regional and so on)
- areas of coverage of the digital preservation programme: transfer of records (including conversion and reformatting); appraisal, selection and acquisition; storage and maintenance; access and dissemination; the implementation of standards; the use of procedures; quality control; and technical infrastructures
- processes for monitoring and reviewing the preservation programme.

Rather than provide a simplified example of a preservation policy here, it is more appropriate to direct readers to the wide range of policies available through the Internet, which can be reviewed and assessed when developing an institution's own digital preservation policy. Some key resources are listed in Figure 4 below.

Developing a preservation policy is a complex process that can address both electronic and manual records. In addition to the citations above, readers will find some of the reference materials in the *Additional Resources* document valuable when considering how to develop a preservation policy.

See the *Additional Resources* tool for more guidance on preservation planning and policy development.

Figure 4: Examples of Preservation Policies

Institution	Policy Example Available at
Australian Digital Recordkeeping Initiative	http://www.adri.gov.au/model-plan.doc
Cornell University Library, United States	http://commondepository.library.cornell.edu/cul-dp-framework.pdf
Digital Preservation Coalition, United Kingdom	http://www.dpconline.org/graphics/handbook/
Digital Preservation Europe (PLATTER) (successor to ERPANET)	http://www.digitalpreservationeurope.eu/platter/
Electronic Resource Preservation and Access Network (ERPANET)	http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf
InterPARES Project (University of British Columbia, Canada)	www.interpares.org/public_documents/ip2(pub)preserver_guidelines_booklet.pdf and http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_appendix_21.pdf
Yale University Library, United States	http://www.library.yale.edu/iac/DPC/final1.html

Preparing a Risk Assessment

Preservation decisions should aim to minimise the risk that electronic records will become inaccessible over a defined period. A risk assessment analyses the dangers that electronic records may become unusable and the impact or consequence of losing the record, such as the risks faced by the organisation or the public if the evidence is not available. An effective risk assessment is based upon analysis of a set of standard criteria, including both generic and specific risk factors. It is important to assess the risk associated with creating and storing electronic records, so that priorities can be established for action and so that unforeseen or emergency situations can be dealt with before they become disasters.

There are common generic risk factors that can potentially affect all digital objects. For instance, a generic risk might be that knowledge about the use of a particular type of format may be widespread, providing unscrupulous individuals with information they could use to access electronic records without authorisation and perhaps change the records, misuse them or expose them to other risks. Another generic risk might be inconsistency in software support, meaning that the organisation may have access to regular technical support for one software package but little or no formal support for another, placing some of the electronic records at risk.

The results of the risk assessment should be used to determine the urgency of any preservation action. If the assessment shows there is a low risk, the next action may be to conduct another assessment at a future date. If the assessment shows that there are areas of high risk, it may be necessary to take immediate action to remove or reduce the threats.

A useful risk assessment tool created by the International Records Management Trust is the *E-Records Readiness Tool,* available through The UK National Archives' website at http://www.nationalarchives.gov.uk/rmcas/documentation/eRecordsReadinessTool_v2_Dec2004.pdf.

Monitoring Technological Change

It is important to monitor technological change to identify potential risks to specific records. For example, if the organisation ceases to receive technical support for a particular software product, the risk increases that the formats supported by that software might become difficult to access and preserve. Similarly, if a particular piece of hardware becomes obsolete but is essential for storing certain digital objects, immediate action will be required to move those objects to a more suitable storage environment, after which an overall assessment of the storage environment should be completed to identify any other such 'single point of failure' dependencies. The act of monitoring such changes in technology and service allows the organisation to maintain high-quality and current preservation strategies and avoid costly data recovery activities.

Assessing Record-keeping Needs

It is also important to assess the current and future record-keeping needs of the organisation and to identify vulnerable and valuable documentary evidence that needs to be preserved. The records survey discussed in Module 2 should provide an overall understanding of the nature and scope of all electronic records within the organisation; from this survey it should be possible to establish priorities for action when implementing a preservation programme.

Once the risk assessment and records assessment have been completed, and any urgent technology concerns have been identified, it is possible to establish priorities for action. For example, some records may have great evidential value and so need to be protected as a priority. Others records may be less critical. However, if a software system is about to change, some records may be at risk during the conversion process. Records that were previously assessed as low risk may need to be reconsidered and given a higher priority for preservation.

Establishing Security and Access Controls

Organisation-wide security controls are needed, such as limitations on user access to computer systems and control over the physical storage of electronic records. A successful programme will have such controls for the management of records throughout the life cycle and in any location. It is critical to ensure that security and access measures are in place when preserving records as authentic evidence for the long term, in order to protect records from unauthorised change or deletion. Although the nature and degree of the controls required will vary considerably depending on the specific needs of the organisation, the following issues will need to be addressed.

- The *physical infrastructure* required to store and manage electronic records must be protected from accidental or deliberate damage. A range of controls may be implemented, including physical access controls, intruder detection systems, fire detection and suppression systems, and backup power supplies.
- *Information technology (computer) systems* should be protected from intrusions by external hackers and other unauthorised users, and from damage caused by malicious code or other forms of software designed to infiltrate or attack a computer system. Counter-measures may include the use of password controls, firewalls and anti-virus software.
- *Access and permissions* must also be controlled. The system must ensure that both internal users and anyone using the system from external locations have appropriate access rights to the stored content. For example only those system users charged with preservation tasks, such as migration, should have the authority to alter or delete stored objects in a digital archival repository, where records are to be kept for their enduring value. Other types of access controls may be required if the content of the digital storage facility is considered sensitive or confidential. Appropriate systems for authenticating and authorising user and system access should be implemented. As a minimum, authentication can be achieved by creating specific operating system user accounts with appropriate permissions and capturing appropriate audit data about access to and and use of a record as part of the record's metadata.

Ensuring the Integrity of the Digital Object

The integrity of a digital object arises from the assurance that it has not been altered in any unauthorised or undocumented manner; the whole purpose of digital preservation is to retain the electronic record as it was originally created and kept. Possible threats to the integrity of an object include accidental corruption, deliberate alteration by an unauthorised user and alteration caused by malicious code, such as a virus. In order to protect the integrity of a digital object, it must be possible to confirm that it has not been altered in any unauthorised manner, such as through corruption, deletion or addition. Clear and consistent processes must be used to monitor the integrity of the content, context and structure of all digital objects, along with their metadata.

Even though many potential threats can be prevented by adopting the security and access measures discussed earlier, it is also essential to test the integrity of the records periodically, to search for any corruption to or other alteration of the data.

Checksums, which have already been mentioned in this module, provide a simple and effective means of checking data integrity. A checksum is created by calculating the binary values (ones and zeros) in a block of data and storing the results with the data. When the data is retrieved, a new checksum is calculated and compared with the existing checksum. A non-match indicates an error.

One approach would be to identify the checksum value for every file to be transferred to the storage repository and compare it to the checksums of the same files after they have been transferred. Any difference will demonstrate that the file has been altered in some way. Exactly what is different will have to be determined upon closer examination, but the checksum process at least makes it possible to see if anything has changed over time.

A number of free tools for generating and comparing checksums are widely available, including Jacksum and JHOVE, which were introduced earlier in this module.

Managing Metadata

Metadata needs to be maintained not just from the time the record was created but also to record any active or passive preservation processes; any physical or logical changes to a digital object; or any other changes to the nature and content of the record. All changes to metadata should themselves be auditable, so that an accurate and up-to-date history of the record can be viewed at any time.

Unlike metadata relevant to the creation and use of records, preservation metadata specifically supports the process of digital preservation. It focuses on

- identifying provenance (who owns or has custody of the digital object)
- confirming the authenticity of the object
- describing the technical environments in which the digital object has been created
- tracking preservation activities
- identifying intellectual property and other rights related to the digital object.

Preservation metadata must be linked to the digital objects they describe. If possible, metadata should be stored in some form of database, to allow easy querying and maintenance. If specialist digital repository software is being used, this software should provide integrated facilities for the management of metadata. In effect, preservation metadata becomes a record in itself, and it is useful to manage it as such.

When considering preservation strategies, it is important to focus on the specific metadata needed to document preservation activities.

> For a broader discussion of the role of metadata in electronic records management, see *Module 1.*

It is fundamental to effective preservation strategies that the relationship between any digital object and its metadata be maintained continually: in other words, the link

between the two should never be broken. This coordination can be achieved in a number of ways, but as mentioned earlier, it is strongly recommended that a persistent, unique identifier (see Unit 4.1) be assigned to every digital object at the time it is brought into the digital repository, and that this identifier be recorded within the associated metadata to provide this persistent and precise link.

PREMIS

In 2003, a working group called PREMIS or PREservation Metadata Implementation Strategies Working Group was established in the United States to identify and define specific metadata elements required to support the preservation of electronic records and digital objects. The group was also tasked with identifying and evaluating strategies for capturing and storing preservation metadata in digital preservation systems. The result is a published data dictionary extending more than 200 pages, which includes definitions of a wide range of different specific metadata elements (referred to in the dictionary as semantic units) that might be captured in order to preserve electronic records or other digital objects.

The importance of such a tool as the PREMIS *Data Dictionary* is that it provides solid guidance for anyone considering establishing effective and reliable strategies for the preservation of electronic records; not only does it include detailed information about the data to be captured – the metadata – but it also offers insights into the steps involved in establishing a digital preservation programme.

Figure 5 shown below illustrates two of the entries in the PREMIS *Data Dictionary.* The examples show how the data dictionary standardises the specific components of the metadata to be gathered, such as how precisely to identify the name of the software application used (1.5.5.1) and how precisely to identify the version of that software (1.5.5.2).

Figure 5: Examples of Entries from the PREMIS *Data Dictionary*

Example 1

Semantic unit	**1.5.5.1 creatingApplicationName**		
Semantic components	**None**		
Definition	A designation for the name of the software program that created the object.		
Data constraint	None		
Object category	**Representation**	**File**	**Bitstream**
Applicability	Not applicable	Applicable	Applicable
Examples		MSWord	
Repeatability		Not repeatable	Not repeatable
Obligation		Optional	Optional
Usage notes	The *creatingApplication* is the application that created the object in the first place, not the application that created the copy written to storage. For example, if a document is created by Microsoft Word and subsequently copied to archive storage by a repository's Ingest program, the *creatingApplication* is Word, not the Ingest program.		

Example 2

Semantic unit	**1.5.5.2 creatingApplicationVersion**		
Semantic components	**None**		
Definition	The version of the software program that created the object.		
Data constraint	None		
Object category	**Representation**	**File**	**Bitstream**
Applicability	Not applicable	Applicable	Applicable
Examples		2000	1.4
Repeatability		Not repeatable	Not repeatable
Obligation		Optional	Optional

For more on PREMIS, see the official website at http://www.oclc.org/research/projects/pmwg/default.htm.

Managing Storage Media

The media on which the records and metadata are stored must be managed and refreshed as required. Part of storage management is concerned with the physical storage of the collection and, in particular, the media on which it is recorded. No computer storage medium can be considered archival, in the sense that it will never need to be superseded because technological obsolescence is inevitable. In many cases, the technologies required to access a certain medium will become obsolete long before the medium itself begins to deteriorate. Therefore, the physical media on which electronic records are stored will have to change over time.

This inevitable process of change is managed through the technique of media refreshment. As discussed earlier, media refreshment involves the periodic transfer of digital information from one storage medium to another which may be of a different type.

It is essential that the storage system be backed up and that multiple copies of all data are stored in order to provide a safeguard: in other words, there should be additional copies available so that if one copy is not usable another can be accessed. It is recommended that three copies of all electronic records be preserved, including one copy stored in a separate geographical location, to prevent loss in the event of disaster. Ideally, at least two types of storage media should be used for the three copies. For example, one copy might be stored on hard drives and the other two on CD disks or tape drives.

This diversity of storage reduces the overall technology dependence of the stored data. If the same type of storage media needs to be used to store multiple copies, then at the very least it is wise to use different brands of the same media storage type, or select items from different batches, to minimise the risk that any data may be lost. One brand or one batch may have manufacturing flaws, for instance, and if all three

copies of data are stored on the same brand, manufactured at the same time, with the same flaw, then the risk of loss is dramatically increased.

Clearly articulated policies are also required for the creation and management of system backups so that all actions taken to preserve electronic records are methodical and well managed. Backup procedures should be fully documented, and the viability of existing backup copies, including the ability to restore the data onto the computer system using the backups, should be tested periodically to make sure all systems and processes are reliable.

Media should be stored and handled in accordance with recommended good practices. The following are important storage guidelines.

- Always store media in the correct cases, and always store them in their containers when not in use.
- Do not leave storage media in computer drives unnecessarily, since prolonged exposure in the computer can cause both heat and mechanical damage.
- Media should not be allowed to come into contact with liquids, dust or smoke, nor should they be exposed to either extreme heat or direct sunlight.
- All media types should be stored vertically, preferably within a locked, fire-resistant safe.
- Magnetic media should be kept away from potential sources of magnetic fields, including electrical equipment.
- Media that has been stored in climate-controlled environments should be left in the operational area for at least 24 hours before they are used; the media need to acclimatise to the changed environment so that they are not adversely affected by the different temperature and relative humidity.
- Computer drives should be maintained and cleaned on a regular basis, in order to prevent damage to media.

It is important to check all stored media regularly: a six-monthly cycle is recommended. The check involves looking for any visual signs of damage to the media or the storage container and checking a random sample electronically to confirm the readability of the data. A checksum, as discussed above, may also be done periodically on selected files to test if there have been any changes to the numerical value, which will signal a possible loss of integrity in the file.

More detailed guidance on handling and storing particular types of media may be found in literature from the relevant manufacturers, or in national and international standards. One example is the British standard BS 4783: 1988: *Storage, Transportation and Maintenance of Media for Use in Data Processing and Information Storage.* ISO has also developed a wide range of standards related to the quality of storage media, including ISO 18923: 2000, *Imaging Materials: Polyester-base Magnetic Tape – Storage Practices* and ISO 18925: 2002: *Imaging Materials: Optical disc Media – Storage Practices.*

See *Additional Resources* for more information on these and other media storage standards and guidelines.

Managing Content

Since a collection of electronic records will always expand as new records are added to the system, it must be possible to ensure the security and protection of records as new material is added, or if records are exported to other locations, updated to capture new metadata or deleted according to established retention schedules and protocols. Procedures need to be established and maintained to ensure that these changes can be made to the contents of a digital storage system without affecting the integrity of the records.

Managing content also involves ensuring that backup copies are retained of all records, in the event of loss in an emergency. A backup copy is not the archival copy but is a secondary copy of the data, kept to replace the original item if required. There is also a difference between backing up data or systems and replicating data, as discussed in the next unit as part of an examination of the concept of trusted digital repositories.

Planning for Emergencies

The digital storage system must be protected against both natural and human-caused disasters. This protection comes from establishing a business continuity plan, which identifies how an operational service will be restored in the event of a major disruption. Policies and procedures need to be established to clarify how the records will be restored in the event of disaster. Ideally, a digital object emergency plan should be closely tied to an organisation-wide business continuity plan, and both plans should be tested periodically, updated as needed and reviewed carefully in the wake of an actual disaster.

A comprehensive business continuity plan should include the following elements.

- Detailed instructions for staff to follow in the event of different types and scales of emergency.
- Contact details for key staff and for any emergency services, including specialists in disaster recovery who may be engaged as contractors.
- Instructions for restoring the content of the digital collection from backup copies.
- A complete description of the hardware and software infrastructure in place to manage the digital objects, with enough information to allow the organisation to acquire replacement equipment or new software if required.
- Copies of crucial documentation related to the preservation process, such as operating procedures and manuals.

The organisation should also have access to copies of all the software required to operate the computer systems. That way, if the original hardware or software fails for

any reason it will be possible to keep the system operational, at least until long-term repairs can be made.

Testing the business continuity plan is every bit as important as developing it. Testing allows staff to rehearse responses and reveals flaws or gaps in the plan. While a full-scale disaster recovery drill can be a major and time-consuming exercise that is only carried out occasionally, individual elements of the plan can and should be tested regularly. For instance, it is useful to contact every person on the contact list periodically, to confirm that they are aware of their role and that their contact details have not changed. Any changes in the organisation – including the expansion of facilities, changes in equipment and technology, or the replacement of staff members – must be reflected in the business continuity plan, so regular reviews and updates are essential.

The basic actions outlined in this unit can help ensure that electronic records are safe and secure within the organisation, at least until a more formal digital preservation programme is established. The next unit looks at the establishment of a trusted digital repository: a formally established facility created to acquire and preserve electronic records according to accepted record-keeping standards.

UNIT 4.3

PRESERVING ELECTRONIC RECORDS IN A TRUSTED DIGITAL REPOSITORY

Establishing an electronic document and records management (ERMS) system is an effective way for an organisation to protect records and ensure their authenticity while they are being created and used within that organisation. However, the long-term security of those records is not necessarily guaranteed by storage in an ERMS. Electronic document and records management systems have endeavoured to address the issue of authenticity through the use of audit trails, but these procedures, while important, do not necessarily ensure the ongoing integrity of the records that have continuing historical, administrative, legal or financial value.

In the last decade or so, efforts have been made to create a more standardised environment for preserving electronic records once they are no longer needed for active use. As mentioned in Module 1, the creation of the Open Archival Information System (OAIS) standard has set the framework for the establishment of what is called a 'trusted digital repository.' The OAIS framework provides archival institutions with specific guidelines for acquiring, describing, maintaining and providing access to electronic records.

The development of these 'trusted digital repositories' is in its infancy, and this unit does not and cannot provide a complete prescription for how to create such a technological and organisational environment. As noted by Australian archivist Adrian Cunningham, discussing the situation at the National Archives of Australia (NAA):

> The NAA now has a fully functioning, secure offline digital repository and is accepting and processing transfers of born digital archival value records from agencies. Nevertheless we regard this work as still at the cottage industry or proof of concept stage. We know we need to be able to perform this work on an industrial scale for billions of records. We also know that we must be able to provide greater support for digital preservation work in those agencies that preserve long term, temporary value (i.e. not archival value), born digital records for a long term, in some cases for as long as 120 years. At this time, we do not have the capacity to perform all this work at this scale, even though we are confident we know how. We need our government to recognise our needs in this area and to fund archival operations for both paper and digital records.[1]

[1] Adrian Cunningham, 'Digital Curation/Digital Archiving: A View from the National Archives of Australia,' *The American Archivist* 71 (Fall/Winter 2008): 539–40.

This unit examines the issues associated with establishing an effective and sustainable trusted digital repository and with preserving electronic records so that they remain authentic and reliable evidence. Part of the solution involves identifying, selecting and installing software designed specifically to preserve electronic records, and this unit provides information about different software packages available. However, if an organisation does not find any of these software packages suitable for the moment, interim solutions need to be found in order to gain control over records and start laying the ground work for the development of a trusted digital repository.

For example, an ERMS can be used to store electronic records; as long as access is restricted and the integrity of the data is checked regularly, it may be possible to protect digital objects securely for a few years. An ERMS should have suitable audit features, which can be used to track access to records and any changes or deletions made. However, relying on regular computer servers is not recommended as adequate storage systems for valuable organisational records and information; the lack of control over metadata, the easy ability to alter or delete records and the difficulty of controlling access are all major concerns. Any organisation wishing to protect its records must consider at some time, preferably sooner rather than later, depositing its records in a trusted digital repository or establishing its own trusted digital repository.

The specific issues addressed in this unit include: the concept of a trusted digital repository, requirements for establishing and maintaining a trusted digital repository, selecting hardware and software solutions, understanding the ingest process and information packages, choosing storage devices, preparing records for preservation in the repository, ingesting records into the repository, destroying original records, monitoring the status of the preservation programme and staying current with technological change.

What is a Trusted Digital Repository?

In its 2002 report on the development of trusted digital repositories, called *Trusted Digital Repositories: Attributes and Responsibilities,* the Research Libraries Group (RLG), a research organisation based in the United States, defined a trusted digital repository as an institution designed to provide long-term access to digital resources.

See *Additional Resources* for more information about the RLG report.

As part of that responsibility, a trusted digital repository must ensure that it

- will maintain digital resources in a long-term and committed manner
- will meet or exceed standards for management, access, and security
- can be audited to ensure appropriate performance and quality management.

In other words a trusted digital repository must ensure the reliability, trustworthiness and accuracy of records; and it must be transparent and accountable to users and

stakeholders while ensuring the long-term preservation of digital archival records. Therefore, a trusted digital repository is not simply a computer program designed to store records; it requires a well-planned and effective administrative, procedural, fiscal and technological infrastructure.

There are two principal types of trusted digital repository: centralised institutions maintained by the creator of the records and decentralised institutions, where the creator of the records transfers custody of those materials to a trusted third party service. Another model that is increasing in popularity is the networked trusted digital repository, where several similar institutions, such as different archival repositories, combine their resources to share responsibility for managing the electronic records of many different creators.

Centralised Management

Under this model, an organisation creates its own trusted digital repository for the preservation of its own electronic records. Examples include a national archival institution establishing a trusted digital repository for the preservation of government records or a university archival facility establishing a repository for official university records. The benefits of such a centralised approach include complete control over the preservation process, limited risk of loss or damage resulting from the involvement of external parties and the ability to establish and oversee a complete life-cycle approach to records management, from creation to final disposal and ongoing preservation. The drawbacks include difficulty sustaining the programme if the organisation is not large enough to commit adequate resources; and the need for high levels of expertise and knowledge in order to achieve success.

Trusted Service Party Provider

Under this model, the organisation – such as a government office or even a national archival institution – engages the services of an external agency, which establishes and maintains the repository on behalf of the organisation. Agencies might include other archival institutions or commercial service providers. It is essential that the service provider has a proven track record for the preservation of authentic and reliable electronic records. The organisation and the service provider need to confirm the conditions and obligations service in a clear and detailed written agreement. Also the service provider needs to provide regular evidence that it continues to meet the standards set by OAIS as well as the requirements of the creating agency.

Networked Repository

Under this model, liked-minded institutions, such as archival facilities, combine resources to promote the preservation of electronic records. One institution may provide the infrastructure and technological support required to ensure the trusted digital repository is operational, and other institutions pay for the storage of their own electronic records. Essentially the intention is to establish a network of archives and maximise resources by equipping one institution with the technological and infrastructure needed, and requiring the other agencies to support the maintenance of the programme through financial and other contributions. Often a networked

repository includes establishing 'mirror sites' where copies of records are maintained, so that copies in one site can be retrieved and used if another site becomes inaccessible.

The networked repository model is considered the most cost effective and practical for the following reasons.

- A networked approach allows smaller institutions to preserve electronic records without having to shoulder the prohibitive infrastructure costs.
- It allows larger institutions to offset some of the costs of establishing their own repository, since smaller institutions will be paying for the storage space they use.
- It allows for continued access to records in the long term, particularly in the event that a smaller institution becomes defunct.
- It can create a facility that provides a good practice example for other similar institutions to follow.

Requirements for a Trusted Digital Repository

As outlined in the RLG report mentioned earlier, several requirements must be met in order to ensure a trusted digital repository is, in fact, trustworthy and sustainable. These requirements are summarised below.

1. The repository needs to be compliant with the guidelines set in the OAIS reference model to meet accepted criteria and standards for operation.
2. The organisation needs to commit the resources and support needed to ensure the repository is administratively and financially sustainable.
3. The technology used must be suitable to the needs of the repository, and all systems, particularly storage and access systems, must be secure.

The various specific requirements that should be in place to maintain a trusted digital repository are discussed in more detail below.

OAIS Requirements and Other Standards and Guidelines

All trusted digital repositories must meet the requirements set out in the Open Archival Information System (OAIS) reference model, which has become the *de facto* standard by which all trusted digital repositories are measured. Conforming to the OAIS model provides the institution with a common framework that allows for consistency of practice. Particularly useful is standardisation of the terminology and concepts used to describe records, as well as the availability of an established and tested model for the capture of preservation metadata. As well, the repository needs to follow good practice in the following operations:

- adherence to established standards for metadata encoding, metadata management and records description
- establishment of an appropriate temperature controlled environment for storage

- creation of backups in an appropriate and timely manner, that meet national or internal standards, if any
- establishment and maintenance of emergency recovery processes, business continuity and contingency planning and other risk mitigation requirements
- establishment of adequate security measures, including hierarchical password access, audit trails, firewalls, virus protection and, if required, encryption.

Administrative and Financial Sustainability

Senior management in the organisation must provide ongoing and committed support to the establishment and maintenance of the repository. This support includes

- providing adequate financial support
- ensuring comprehensive policies and procedures are in place to allow the operation to work effectively and to be open, transparent and accountable, including policies and procedures for the migration of records, maintenance of records systems, disaster preparedness, data recovery, audits and the management of security systems
- establishing and supporting clearly defined roles and responsibilities among staff for various tasks within the repository
- ensuring that the repository has appropriately trained and adequately compensated staff, maintains proper staffing levels and supports ongoing professional development
- establishing formal succession plans for key staff
- establishing contingency plans in the event of any change in service, including a decrease in or cessation of operations.

Technological Suitability and Security

The technological requirements of a trusted digital repository should meet or exceed those outlined in the OAIS model and other records standards, including the following.

- A range of preservation strategies should be built into the system, such as ensuring additional copies are available, establishing mirror sites, or ensuring the stability of electricity and other technologies required to maintain the system.
- Plans should be developed for the regular and timely upgrading and replacement of hardware or software, including ensuring staff are trained and able to carry out the upgrades.
- The systems should be audited regularly for their technical quality and the effectiveness of operations.
- Storage systems should be flexible (in that they can accommodate technological changes) and scalable (so that they can be expanded as needed).
- Appropriate security features need to be in place, including: firewalls, server encryption, audit trails, checksums, passwords, backup systems and others as appropriate.

Legal and Organisational Framework

When planning a trusted digital repository, it is also important to consider other requirements, including legal requirements that may impinge upon records preservation: examples include evidence laws; copyright acts; national archives and records legislation; and privacy laws.

Selecting Hardware and Software Solutions

After determining what type of digital repository will be developed and creating the mechanisms for ensuring that the repository meets appropriate standards of practice, the next step is to decide which of many hardware and software solutions will be most suitable for the repository to use. There is no one 'ideal' hardware or software solution for a digital repository. Ultimately, it is up to each organisation to examine different options and select one that best suits its needs.

Below is a short overview of six software options current in use around the world. The first five – DAITSS, DSpace, Fedora, Greenstone, and LOCKSS – are freely available, open-source packages available for download on the Internet, while the sixth – CONTENTdm – is a proprietary software program that must be purchased. (The packages are presented in alphabetical order, not in any order of preference.)

As will be seen from the summaries provided below, none of the existing digital repository software programs available today offers a complete solution to ingesting, preserving and accessing electronic records.

Since most of these software applications are open source and freely available they do not always add direct costs to the digital repository's budget. However, maintaining such programs requires expertise, and there can be costs associated with obtaining new releases or updates to 'free' software, so users must be careful about the direct and indirect costs involved. There may be other indirect costs, including hardware and staff and training time involved with developing and implementing the solution.

CONTENTdm

CONTENTdm is a commercial, proprietary software program, designed to ingest batches of electronic records from specific user groups. Once the records have been submitted, an administrator can modify the metadata and preserve the records in a digital repository. The software is interoperable with the Online Computer Library Center, which can host the digital repository and make it available via WorldCat, which claims to be the world's largest network of library content and services. CONTENTdm can also store and display any digital object format. The system has limited preservation functionalities and cannot support METS files. Users interested in this software should research their preservation, metadata and auditing requirements carefully before committing to it.

For more information about CONTENTdm, see the official website at www.contentdm.com. For more on WorldCat, see the official website at http://www.worldcat.org.

DAITSS

DAITSS, which stands for 'Dark Archives in the Sunshine State,' was developed by the Florida Digital Archive and Florida State University. The software is designed to support the creation of a digital preservation repository, and functions include ingesting records (including data migration) and managing and disseminating information. In keeping with its name, 'dark archive,' the software is not intended to allow public access, in order to ensure the best security possible. Therefore, the software contains no public interface, but it can be used in conjunction with other access-oriented software. However, in 2008, a DAITSS 2 project was initiated to rewrite the DAITSS application for web-based use, while keeping the same functionality as DAITSS. The software is sophisticated but not immediately user friendly, so implementation requires strong support from information technology specialists; readers interested in this software should monitor the progress of DAITTS 2 in the event that it proves easier to install.

For more information about DAITSS, see the official website at http://daitss.fcla.edu/.

DSpace

DSpace was developed by Massachusetts Institute of Technology (MIT) and Hewlett-Packard in order to allow users to deposit digital objects into a repository using a web-based interface. The program was initially created to allow MIT faculty and researchers to deposit their personal papers in the archival facility. The system itself is highly library oriented, emphasising, for example, item-level descriptions of materials. The software also does not control the creation of authorities, such as names or other index terms, making it difficult to obtain accurate results when searching for records. However, organisations that decide to use DSpace can develop their own authority controls, improving the success of searches.

Readers will want to refer to the resources in the *Additional Resources* tool for information about establishing authority controls.

The software can accept multiple file formats but only some of them can be converted to archival formats, and the software only supports Dublin Core metadata sets, limiting some of its functionality. A fair amount of technical knowledge is required to install and implement DSpace, and so access to technical specialists will be necessary during the development of the software.

> For more information about DSpace, see the official website at www.dspace.org.

Fedora

Fedora stands for 'Flexible Extensible Digital Object Repository Architecture.' This software package was jointly developed by Cornell University and Virginia University Library and then tested by Yale University and Tufts University. Fedora is designed as a digital storage software system and can both preserve individual digital objects and maintain relationships among different digital components that make up a complex electronic record. All objects and related metadata are stored as .xml files, which has become the *de facto* encoding standard for the preservation of electronic records.

> For more information about Fedora, see the official website at http://www.fedora-commons.org/.

Greenstone

Greenstone was developed in the mid-1990s as part of the New Zealand Digital Library project, based out of the University of Waikato. The software was designed to serve as a tool for building digital libraries, and so it is not specifically focused on preserving archival records as evidence. The package was distributed by the University in coordination with UNESCO. The program is designed to be easy to use and users do not need to be technical experts. The software allows users to customise metadata sets to suit the specific needs of the repository, and the program also allows records to be exported to external storage devices, such as USB drives or removable hard drives, making it easier to provide access. Greenstone operates with a web interface, making it easy to use, but since it has not yet developed a dedicated preservation component, anyone using it for ongoing preservation of records as evidence needs to establish rigorous procedures for controlling security and access to items stored in the repository.

> For more information about Greenstone, see the official website at www.greenstone.org.

LOCKSS

LOCKSS means 'Lots of Copies Keeps Stuff Safe.' The software, which was developed by Stanford University Libraries, was initially designed to enable libraries to preserve web content and electronic publications through automatic capture of web resources. The idea was to support redundancy or the preservation of multiple copies of electronic materials, and there seems to be an emphasis on the management of web-based publications rather than electronic records that need to be preserved as evidence. The software conforms to the OAIS model in terms of storage and access and is actively used by libraries and publishers. The programmers are currently working on a feature that would perform migration on demand as well as migration on access, meaning that a record would be migrated once it is opened or created by a user the first time. The system also automatically compares multiple copies of a document to detect and repair any errors.

> For more information about LOCKSS, see the official website at www.lockss.org/lockss/Home.

It is important to remember that any software solution needs to be tested, more than once, using duplicate copies of records, before the software is formally implemented. Do not test software programs using original records, or any problems may very well result in a loss of authentic evidence.

The Ingest Process and Information Packages

For every transfer of records that a system receives (referred to as a Submission Information Package or SIP in the OAIS model), the system produces two outputs: an Archival Information Package (AIP) and a Dissemination Information Package (DIP) which are managed in the trusted digital repository.

- The SIP contains the digital object itself, as transferred from the creator to the archives. The SIP also contains basic metadata regarding the creator, writer and addressee, in order to identify the provenance of the object.
- To the SIP is added virtually all the information about a record available through the different metadata sets, to create the AIP. The information that is added includes, for example, information about finding aids, authority controls, migrations, checksums and data recovery information.
- The DIP is a pared down version of the AIP, containing the digital object along with enough metadata needed to identify the provenance of the record and identify it for research purposes.

It is important for the records professional involved in establishing a trusted digital repository to understand the general nature of, and importance of, these different information packages, since they contain the data and instructions needed to convert the actual bits and bytes of data received in the digital repository into meaningful information and records.

Choosing Storage Devices

Once the software is chosen, a subsequent task is to determine the best storage for protecting the digital objects. The software chosen, whether Greenstone or Fedora or any of the other packages discussed earlier, does not dictate what hardware to use to store and backup electronic records. Research has shown that the best storage structure for a digital repository is called a SAN or Storage Area Network. A SAN is best described as a network of storage devices that can be managed independently of the general purpose network to which it is connected.

In order to maximise the amount of time electronic records can be stored on a particular storage medium, it is important to select media that can meet the following conditions, as identified by digital preservation expert Adrian Brown from the National Archives of the UK.

1. The media should have a life span of at least 10 years. It is not really a benefit to choose media with a longer life span since it is likely that the technology needed to use the media may be obsolete by then.
2. The media should have enough capacity to hold the records required; the fewer separate storage devices needed, the easier it will be to manage the technology.
3. The media should be well tested, so that it is known to cause few or no errors reading or writing data, and it should be 'read only,' which means the media cannot be reused by accident once data have been recorded the first time.
4. The media chosen, along with any necessary hardware and software, should be proven technology, well known in the industry and widely available.
5. When assessing the cost of the storage media, consider both the cost of the technology itself and the ongoing costs of owning and managing the technology, including the initial purchase and any anticipated maintenance, storage, repair and support costs.
6. The media should be tolerant of changing environmental conditions and, as much as possible, should not be vulnerable to physical damage, such as accidental erasure or damage from magnetic fields.

See *Additional Resources* for more information about Adrian Brown's 2003 guidance note *Selecting Storage Media for Long-Term Preservation.*

As noted earlier, a trusted digital repository should also establish policies with regard to backups, offsite storage and other procedures required to protect electronic records. And the organisation should factor into its budget any anticipated costs for hardware replacement, data refreshment, replication of records and migration of files.

Preparing Records for Preservation

Normally, preservation actions are applied to specific classes or series of records, not to every record in an organisation's server or to individual records, except in special circumstances. This approach is premised on the assumption that a strong and effective classification scheme and retention and disposal schedules have been established, making it possible to identify and capture records at the series level.

Ensuring the Authenticity of Incoming Records

As has been emphasised throughout this training programme, the evidential value of electronic records rests heavily on the fact that they can be considered authentic and reliable: that the records are what they purport to be. Before records are transferred to any digital storage environment, therefore, it is essential to confirm that they are accurate, authentic and reliable. To do so, it is important to confirm that any metadata needed to contextualise them is, in fact, in place and associated with the record.

Usually, records that have been created and managed in a controlled environment, such as an ERMS, can be considered authentic because the ERMS software is designed to capture essential metadata about when the record was created, by whom, how it was used, and so on. But if records were created outside of an ERMS, it is highly likely that there is inadequate metadata attached to the document and no audit trail to prove that the record is authentic.

As a useful analogy, it may be helpful to imagine that an archival institution has received a shoebox of photographs on the doorstep, with no accompanying information provided. The archivist would have to carry out extensive research to identify and describe the photographs, and even then it may not be possible to prove that the people in the images were in fact the people the archivist identified. The photographs may be interesting, and they may be informative, but they cannot serve as evidence of the actions and lives of the people who created them or are shown in them.

The same scenario is true for uncontrolled records. The records may well have long term value to the organisation, but if they lack sufficient metadata, such as information about the creator, writer, addressee or dates of creation, they cannot be considered authentic. It may be possible to confirm their authenticity if the records manager is able to add information confirming the context of their creation and the manner in which they were kept prior to ingestion, but obviously the effort and energy are significant. The organisation will be better served by establishing clear and effective record-keeping procedures, including capturing metadata for electronic records as a matter of course, before embarking on digital preservation as a means of protecting an authentic electronic record.

So what information is needed to confirm a record is authentic before it is transferred to digital storage? Below are three issues that need to be considered when striving to ensure records are authentic.

- The records should have been created and maintained in a controlled environment, with unbroken custody.

- There should be adequate security controls in place to ensure records have not been altered or damaged.
- The record should not change once it is ingested into the system.

These three examples represent only a small sample of the minimum requirements for preserving authentic records. For specific guidance on minimum requirements for the preservation of authentic electronic records, see for instance the InterPARES baseline requirements for confirming the authenticity of electronic records.

> The InterPARES baseline requirements are available at http://www.interpares.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf.

The following metadata about the record need to be captured as the record is moved into the digital storage repository:

- the date the record was ingested
- the name of the person responsible for the transfer
- the effect of the ingest process (if any) on the content, context and structure of the record or on the ability to access it
- any variations between the source record and the ingested record that need to be known in order to confirm that the ingested record can be considered authentic evidence
- any descriptive information about the context of the records creation and use – including archival descriptions of the larger group of records within which the individual record fits – in order to document the provenance of the record.

Ingesting Records into the Digital Repository

Records should only be transferred into the repository after the 'ingest' process has been successfully tested. In order to ensure records are transferred to storage successfully, the following ingest procedures also need to be followed.

1. Ensure that each digital object to be transferred has a unique persistent identifier (as discussed earlier).
2. Scan all objects for viruses and other forms of malicious code. (This is one of many reasons why it is essential to ensure that antivirus software is available and up to date.) Ideally, objects should be quarantined for one month after scanning, and rescanned at the end of this period to ensure that very recent viruses can be detected. Any PCs or servers used for the transfer of electronic records should also be protected with up-to-date antivirus software.
3. Before transferring any records, make backup copies of them, verify their integrity and store them in a secure area. These duplicate source records should be held until it is known that the preservation process has been

successful; they may be needed as master copies should something go wrong with the ingest process.

4 Once records have been ingested, test the preserved records again to ensure that any reduction in functionality, or loss of content, structure or format, is within acceptable limits. If the transfer process does not include any normalisation or other steps that affect the file encoding of the digital components, then one means of validating records is to perform a checksum, as discussed earlier. The checksum is run before and after the records are transferred in order to confirm that the records are not altered during the transfer. If the records have been corrupted or altered in any way, the checksum will tag the digital object as faulty.

5 The integrity of all relevant metadata associated with the preserved records should also be verified. In other words, it is important to ensure none of the metadata has changed during the transfer of the records. Metadata should also be updated to record the work that has been done to ingest the records into the repository. If the integrity of the records cannot be verified, the preservation process will need to be repeated on new duplicates of the source records. If at this point the ingest process still results in unacceptable errors, the entire preservation strategy may need to be re-evaluated.

Destroying Source Records

Once the ingest process has proved reliable and the records have been transferred and are safely stored, the organisation can consider destroying the original records. Any decision to destroy records should be based on the organisation's retention and disposal schedules. While keeping authorised copies of records is an important part of the preservation process, keeping 'extras' that fall outside of the framework of retention and disposal requirements might be a violation of the requirements the organisation has set for itself. In the event that those duplicates are accessed or used inappropriately, the organisation could suffer significantly. If the organisation has decided that its preservation programme, as embodied in its trusted digital repository, is in fact trustworthy, then the retention of additional copies of records should be consistent with the organisation's overall preservation strategy and be reflected throughout the preservation system, such as through the inclusion of LOCKSS software (see above).

Monitoring the Status of the Preservation Programme

The integrity of the preserved records – including their functionality, structure, content, context and associated metadata – should be monitored periodically following preservation to ensure the stability of the preserved records and to identify when subsequent preservation treatments are required. If at any time, it appears that any records have lost significant integrity or authenticity as a result of the preservation process, the organisation should immediately stop adding records to the system and investigate the problems.

Staying Current

Any organisation undertaking the implementation of a digital repository must remember to stay current with new technological trends and with changes in standards. The electronic records environment is a dynamic field, and records professionals need to monitor changes in technology and approaches regularly. Regardless of the systems chosen, important qualities of a successful digital repository include

- using open source software whenever possible
- ensuring redundancy (extra copies of records) in case records are corrupted and need to be restored
- choosing scalable and flexible storage architectures so that the organisation can easily support change and accommodate more electronic records over time
- verifying the authenticity of records by running periodic checksums on the records.

After all, the ultimate goal of a trusted digital repository is to preserve documents and ensure the production of authentic facsimiles whenever needed for administrative, fiscal, historical or legal reasons; following appropriate standards and best practices as established by the national and international record-keeping community will support the achievement of that goal.

The next unit discusses some of the current work underway in electronic records research and considers some possible future directions for digital preservation.

UNIT 4.4

CURRENT RESEARCH AND FUTURE DIRECTIONS

Digital preservation is a rapidly evolving field, with new tools and techniques continually being developed. This unit highlights some of the key areas where major new developments may be expected over the next few years. It is important to stay up to date with research and collaborative projects, as they are usually the source for any major changes in practice and can also provide an environment for learning about new initiatives without risking the day-to-day operations of the organisation itself.

This unit contains a brief overview of different electronic records research projects, in addition to the various initiatives already discussed in this module. These different projects are identified by country or region; the goal is to provide a sense of some of the work underway in the field of digital preservation. It is recognised that this list is not comprehensive, and that information about these programs will almost certainly be out of date as soon as this module is made available. However, it is hoped this brief summary will provide readers with some information about a range of initiatives that they can investigate more fully, and will also encourage those interested to follow the progress of some important current research projects.

Australia

As mentioned earlier, an Australian initiative is **XENA,** or 'Xml Electronic Normalizing for Archives,' developed by the National Archives of Australia to help in the long-term preservation of electronic records created by the Australian government.

> For more information on XENA, see the official website at http://xena.sourceforge.net/.

Canada

InterPARES (the International Research on Permanent Authentic Records in Electronic Systems) is a multi-year project designed to research issues related to the long-term preservation of electronic records. Specific topics under consideration include

- the theory and methods related to preserving the authenticity of records created and/or maintained in databases and document management systems (InterPARES 1: 1999-2001)
- researching the reliability and authenticity of electronic records produced in the course of artistic, scientific and e-government activities (InterPARES 2: 2002-2006)
- putting research findings into practice (InterPARES 3: 2007-2012).

For more information, see the official website at http://www.interpares.org.

European Union

CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) intends to define the methodology and infrastructure for digital preservation in Europe. Specific goals include

- building a preservation environment based on the OAIS reference model
- establishing the capacity to preserve the digital resources of a variety of user communities
- incorporating state-of-the art technologies into digital preservation
- developing sustainable digital technologies.

For more information, see the official website at http://www.casparpreserves.eu/.

Digital Preservation Europe (DPE) aims to support collaboration between different European agencies in the development of new approaches to the preservation of digital materials. Among the aims of the agency are

- to raise awareness about the importance of digital preservation
- to coordinate work across Europe in the development of digital preservation strategies
- to develop auditable standards for digital preservation
- to facilitate skills and knowledge through training, coordination, exchange and increased awareness of digital preservation issues and approaches.

For more information, see the official website at http://www.digitalpreservationeurope.eu/.

The **DLM FORUM** is a community of public archival institutions and other agencies interested in a range of issues associated with records, archives and information management throughout the European Union. The DLM Forum developed the *MoReq* (*Model Requirements for the Management of Electronic Records*) standard (European Commission, 2001), which established a common set of functional requirements for Electronic Records Management Systems. A revised version, *MoReq2*, includes requirements for long-term preservation.

> For more information, see the official website at http://dlmforum.typepad.com/.

PLANETS (Preservation and Long-term Access through Networked Services), which is funded by the European Union and involves sixteen partners throughout Europe, aims to deliver a sustainable framework to enable long-term preservation of digital content, increasing Europe's ability to ensure access in perpetuity to its digital information. The project aims to deliver

- preservation planning services to allow organisations to define, evaluate and carry out preservation activities
- methodologies, tools and services for the characterisation of digital objects
- tools to help emulate obsolete digital assets so that they may be accessed and preserved
- a distributed service network to allow for the integration of tools and services
- a test environment to support the development of a consistent body of evidence to support the evaluation of different tools and services
- a dissemination programme to support the use of the tools.

> For more information, see the official website at http://www.planets-project.eu/about/.

New Zealand

As mentioned earlier, the National Library of New Zealand has developed the **NLNZ Metadata Extractor** to extract metadata from different file formats, including PDF documents, image files, sound files, Microsoft Office documents and others.

> More information about NLNZ Metadata Extractor can be found at the official website at http://meta-extractor.sourceforge.net/.

Sweden

ABM Centrum, the coordinating office for Archives, Libraries and Museums in Sweden, was initiated in 2004 as a joint initiative of the Royal Library, National Library of Sweden, National Museum, National Heritage Board, Swedish National Archives and Council for Cultural Affairs. Particular emphasis is given to developing harmonised digitisation strategies and approaches.

For more information, see the official website at http://www.abm-centrum.se/Eng/indexEng.asp.

United Kingdom

The **Digital Preservation Coalition** aims to 'secure the preservation of digital resources in the UK' and is made up of almost 30 government agencies, academic institutions, research agencies and professional consortia. The long-term goals of the Coalition include

- producing and disseminating information on current research and practice in digital preservation
- coordinating efforts to raise awareness among key stakeholders of the importance of sustainable digital preservation, and to secure financial support for national initiatives in the preservation of digital resources
- providing a forum for the development and co-ordination of digital preservation strategies in the UK and internationally
- promoting and developing services, technology and standards for digital preservation
- developing strategic alliances nationally and internationally to address challenges in digital preservation.

The Coalition also publishes a number of reports on significant topics, including the influential *Preservation Management of Digital Materials: A Handbook* and a series of Technology Watch reports.

For more information, see the official website at http://www.dpconline.org/graphics/. The handbook can be accessed online at http://www.dpconline.org/graphics/handbook/ and the *Technology Watch* Reports are available at http://www.dpconline.org/graphics/reports/index.html#techwatch.

DROID and PRONOM

As discussed earlier in this module, DROID (Digital Record Object Identification) and PRONOM are two UK-based initiatives designed to support the identification and management of digital objects.

> More information about DROID can be found at the official website at http://droid.sourceforge.net/wiki/index.php/Introduction and more about PRONOM can be found at the official website at http://www.nationalarchives.gov.uk/pronom.

United States

CAMiLEON, which stands for Creative Archiving at Michigan & Leeds: Emulating the Old on the New, was a joint project of the Universities of Michigan (USA) and Leeds (UK), which finished in 2002-2003. The project sought to develop and evaluate a range of technical strategies for the long term preservation of digital materials. The project's objectives included

- 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 long-term strategy for long-term preservation and access to digital objects
- to consider where and how emulation fits into a suite of digital preservation strategies.

> For more information, see the official website at http://www.si.umich.edu/CAMILEON/.

The **Digital Library Federation** (DLF) includes more than 36 members, mostly academic libraries along with a small number of public libraries, which seek to provide opportunities to research and share experiences about the management of electronic resources. The DLF undertakes a number of different initiatives, including research into

- digital library structures, standards, preservation and use
- archival preservation of electronic journals
- developing online teaching resources
- developing Internet services to expand access to resources.

> For more information, see the official website at http://www.diglib.org/.

The **Electronic Records Archive** is an initiative of the United States National Archives and Records Administration (NARA) to develop methodologies to preserve valuable electronic records of the US government, to ensure long-term access to those records and to establish a government-wide framework for the management of electronic records throughout the life cycle.

> For more information, see the official website at http://www.archives.gov/era/.

The **Florida Digital Archive** strives to provide cost-effective, long-term preservation options for digital materials needed to support research and scholarship in the state of Florida. The repository itself is based on DAITSS, the open-source preservation management software discussed in Unit 4.3.

> For more information, see the official website at http://www.fcla.edu/digitalArchive/.

The **Global Digital Format Registry** (GDFR) project, being led by Harvard University, plans to develop a network of format registries, which any organisation will be able to use to develop preservation plans. The goal is to create software systems that will enable sustainable distributed services to store, discover and deliver representation information about digital formats.

> For more information, see the official website at http://hul.harvard.edu/gdfr/.

RLG (the Research Libraries Group) was established in 1974 as a US-based consortium of libraries; among the projects it developed included RLIN (Research Libraries Information Network), the development of the certification scheme for trusted digital repositories mentioned earlier (developed in conjunction with NARA), and, in conjunction with the Online Computer Library Center or OCLC, the preservation metadata standard PREMIS discussed earlier. In June 2006, RLG merged with OCLC, a non-profit, membership-based, computer library service and research organisation dedicated to increasing access to information around the world.

> For more information, see the official OCLC website at http://www.oclc.org/ca/en/par/default.htm.

The Washington State Digital Archives based in Spokane, Washington, is the nation's first archival facility dedicated specifically to the preservation of electronic records created by public agencies at both state and local levels. The facility, located on the campus of Eastern Washington University campus, was designed as a purpose-built digital repository, with custom designed web interfaces and database storage facilities.

> For more information, see the official website for the Washington State Digital Archives at http://www.digitalarchives.wa.gov/.

Developments in Data Storage

The ability to store and manage ever-increasing volumes of digital data in a cost-effective manner is a major concern for institutions around the world. However, the capacity of storage devices continues to increase as costs decrease, and it is assumed that new and more efficient types of storage technology will continue to reach the market.

The storage capacities of established technologies, such as hard disk and magnetic tape, are continuing to increase significantly. In 2007, single hard disks with one terabyte capacities were available. New technologies such as perpendicular recording, where data are recorded in three-dimensional columns rather than on the two-dimensional disk surface, are expected to increase capacities tenfold over the next few years.

Magnetic tape remains the most common choice for very large data volumes. To illustrate the progression of storage capacities, one of the most widely used tape formats is **Linear Tape Open** (LTO). The first generation of LTO cartridge, released in 1999, could store 100 gigabytes of uncompressed data. By 2002, the second generation offered a capacity of 200 gigabytes. The current LTO 4 format can store 800 gigabytes per tape, and increased capacity increases are expected in future. The issue of data storage technologies, such as data tapes, is complex, and readers are advised to review some of the current resources available for the most up-to-date information on the topic.

> For more information on the LTO program, see the official website at www.lto.org.

Tools called 'flash solid state memory' devices are also becoming increasingly widespread. Although traditionally confined to portable devices such as music players and removable USB storage devices – called 'memory sticks' – the first flash-based hard drives are now available. New forms of storage technology, such as holographic storage, are likely to emerge, offering even greater storage capacity.

However, the development of storage devices is being matched by the dramatic growth in the volume of electronic records. Volumes of data are now being talked about not in kilobytes or megabytes but in petabytes, which are equal to one quadrillion bytes or 1000 terabytes. In the scientific sector, data requirements may reach Exabyte levels – equivalent to one quintillion bytes. In the commercial world, the requirement to retain emails for regulatory compliance is generating massive email archives. The pharmaceutical and petroleum exploration industries also create huge data volumes which require management. For example, a single seismic survey can produce one petabyte, or 1024 terabytes, of data. It is anticipated that the pressure brought by the needs of these sectors will drive research into the development of even more robust data storage solutions. For example, the San Diego Supercomputing Center is developing tools for grid-based storage and preservation. The concept of a grid-based storage facility is that many different information technology resources can be networked together to create one virtual computer storage system. The goal is to distribute the resources required for that storage facility among different IT systems, allowing for improved efficiency and security.

It is expected that future research will result in an increasing number of practical tools and services to support operational digital preservation programmes. New tools and techniques may be expected in the following areas.

- Sophisticated tools to characterise digital objects, enabling an increasingly rich set of properties to be captured automatically. New techniques for semantic analysis and automatic classification of content may allow much cataloguing to be performed automatically.
- Support for rigorous preservation planning, including online technology watch services and advanced risk assessment methodologies.
- The emergence of a global network of registries to support long-term preservation. Such registries are likely to focus on file formats in the first instance but should expand to cover the full range of representation information.
- 'Archival quality' migration tools, and international benchmarking of tried-and-tested migration pathways for common file formats.
- Improved emulation tools and a better understanding of their viability and role as part of a comprehensive preservation toolkit.
- The development of new preservation strategies designed to both meet the challenges of, and take advantage of, developing information technologies.

As well, research continues by standard and policy-setting agencies such as ISO into the standards and requirements for electronic records storage and management; new developments in those areas will have a continuing impact on the management of electronic records.

Staying current with these changing technologies is important to allow the records professional to play a central role in the preservation of electronic records.

STUDY QUESTIONS

The following questions are designed to encourage readers of this module to examine some of the issues raised in more detail and to consider how the general information presented here applies to the specific environment in which these records professionals are working.

1 Define an electronic record element and explain the idea that all electronic records may consist of at least one element or component.

2 Explain the idea of 'characterisation' and explain how it supports the preservation of electronic records.

3 Explain the concept of validating electronic records. What kind of question is being answered by the process of validation?

4 Name two different kinds of software used to characterise digital objects. What is the purpose of these kinds of software programs?

5 What is a persistent identifier and why is it important concept in digital preservation?

6 What is the difference between passive and active preservation?

7 Explain the concept of refreshing data. What concerns or limitations need to be considered when refreshing data?

8 Explain the concept of replicating data. How is replication different from backing up data? What is a major drawback to replication?

9 Explain the concept of emulation and describe three different types of emulation. What are the drawbacks to emulation?

10 Explain the concept of migration. Explain the key differences between emulation and migration and the issues to consider when selecting one preservation strategy over another.

11 What does 'migration by normalisation' mean? What are the strengths of using this type of migration strategy? What are the drawbacks?

12 What is 'migration at obsolescence'? Name at last two advantages and two disadvantages to following this strategy for migration.

13 What is 'migration on demand'? What are the possible benefits of following this strategy, and what are the possible drawbacks?

14 What is the purpose of a preservation policy? What specific issues need to be addressed in a preservation policy?

15 Why is it important to conduct a risk assessment when planning a preservation programme?

16 What security and access controls need to be established to protect electronic records?

17 Explain the concept of the integrity of electronic records. Why is ensuring the integrity of electronic records such an important component of planning a preservation programme.

18 What is a checksum and how can it be used in the preservation of electronic records?

19 Explain the value of managing metadata specifically in relation to the preservation of electronic records.

20 Why is it important to refresh the media on which records and metadata are stored?

21 Describe at least five guidelines for handling the storage media used to manage and preserve electronic records.

22 What is involved with managing the content of electronic records?

23 Why is an emergency or business continuity plan important to preserving electronic records?

24 What elements should be included in a business continuity plan.

25 What is a trusted digital repository? Identify the three main requirements for establishing a trusted digital repository.

26 There are several open source and commercial software products available to store electronic records within a digital repository. Identify at least two different software programs and discuss some of the qualities of different software programs that must be assessed when deciding which software program to select.

27 Explain the difference between the three different information packages that are created during the preservation of electronic records: SIP or submission information package; AIP or archival information packages and DIP or dissemination information package.

28 Name at least five issues to consider when selecting the media on which to store electronic records as part of a preservation programme.

29 What information needs to be captured in order to confirm a record is authentic before it is transferred to digital storage?

30 What metadata need to be captured as an electronic record is moved into a digital storage repository?

31 Explain the five steps involved in ingesting records into a digital repository.

32 What issues need to be considered when transferring electronic records to storage in a digital repository, before any original records should be destroyed?

33 Describe at least three research projects underway today to study issues related to the preservation of electronic records. Which of the many projects described in this module sound like they have the most relevance to the issues of concern within your particular organisational context?

International Records Management Trust
4th Floor
7 Hatton Garden
London EC1N 8AD UK

Phone +44 (0) 20 7831 4101
Fax +44 (0) 20 7831 6303
email info@irmt.org
www.irmt.org

Registered Charity Number 1068975
VAT Registration Number 564 4173 37
Company Limited by Guarantee, registered in England Number 3477376

Produced with funding from the UK Department for International Development