Digital Records Pathways: Topics in Digital Preservation

Module 7: Management and Preservation of Records in Web Environments

InterPARES / ICA
DRAFT July 2012

Table of Contents

Digital Records Pathways: Topics in Digital Preservation

1   Preface

Digital Records Pathways: Topics in Digital Preservation is an educational initiative developed jointly by the International Council on Archives (ICA) and the International Research on Permanent Authentic Records in Electronic Systems Project (InterPARES). It offers training to archivists and records professionals in the creation, management and preservation of authentic, reliable and usable digital records. The program assumes that the user has a solid grounding in basic concepts of records management and archival theory, and builds on that knowledge.

Consisting of eight independent modules, *Digital Records Pathways* addresses the theoretical and practical knowledge needed to establish the framework, governance structure and systems required to manage and preserve digital records throughout the records' lifecycle.. Each module addresses a specific topic of relevance to the management and preservation of digital records. The program is provided free of charge on the ICA website at www.ica.org/.

1.1   About the ICA and InterPARES

The ICA and InterPARES are committed to establishing educational materials for the continuing education of archivists and records managers, to build upon foundational knowledge, disseminate new findings, and to equip archivists and records professionals with the necessary specialized knowledge and competencies to manage and preserve digital records.

The International Council on Archives (ICA) (www.ica.org) is dedicated to the effective management of records and the preservation, care and use of the world's archival heritage through its representation of records and archives professionals across the globe. Archives are an immense resource. They are the documentary by-product of human activity and as such an irreplaceable witness to past events, underpinning democracy, the identity of individuals and communities, and human rights. But they are also fragile and vulnerable. The ICA strives to protect and ensure access to archives through advocacy, setting standards, professional development, and enabling dialogue between archivists, policy makers, creators and users of archives.

The ICA is a neutral, non-governmental organization, funded by its membership, which operates through the activities of that diverse membership. For over sixty years ICA has united archival institutions and practitioners across the globe to advocate for good archival management and the physical protection of recorded heritage, to produce reputable standards and best practices, and to encourage dialogue, exchange, and transmission of this knowledge and expertise across national borders. With approximately 1500 members in 195 countries and territories the Council's ethos is to harness the cultural diversity of its membership to deliver effective solutions and a flexible, imaginative profession.

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) (www.interpares.org) aims to develop the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and provide the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity. The InterPARES project has developed in three phases:

InterPARES 1 (1999-2001) focused on the development of theory and methods ensuring the preservation of the authenticity of records created and/or maintained in databases and document management systems in the course of administrative activities. Its findings present the perspective of the records preserver.

InterPARES 2 (2002-2007) continued to research issues of authenticity, and examined the issues of reliability and accuracy during the entire lifecycle of records, from creation to permanent preservation. It focused on records produced in dynamic and interactive digital environments in the course of artistic, scientific and governmental activities.

InterPARES 3 (2007-2012) built upon the findings of InterPARES 1 and 2, as well as other digital preservation projects worldwide. It put theory into practice, working with archives and archival / records units within organisations of limited financial and / or human resources to implement sound records management and preservation programs.

1.2   Audience

The audience for this program includes archivists and records and information professionals interested in expanding their competencies in the management of digital records. Taken as a whole, the modules form a suite of resource materials for continuing professional education with particular focus on issues influencing the preservation of reliable, accurate and authentic digital records.

1.3   How to Use the Modules

Each module consists of theoretical and methodological knowledge and its practical application, illustrated through case studies and model scenarios. While the modules have been developed by InterPARES Team Canada, and are therefore illustrated with examples from the Canadian context, each module is customizable for a specific domain or juridical context. For wider applicability, they have been translated into the languages of the ICA partners.

The modules can be studied individually according to need and interest, or as a set, covering the range of competencies required. They can be self-administered by individuals, or offered through professional associations or workplace training. The modules also contain a number of templates that allow universities and professional associations to adapt and to develop specific course curricula, on-site training materials for students and professionals on digital recordkeeping and preservation issues. Universities and professional associations are free to adapt the materials and develop their own context-specific course curricula and training kits.

1.4   Objectives

The modules have the following objectives:

- To provide educational resources based on cutting edge research in digital records issues to professional archival and records management associations for the benefit of their members;
- To provide archivists and records managers with the necessary theoretical knowledge as well as procedural and strategic skills to develop, implement and monitor a digital recordkeeping and/or a preservation program;
- To illuminate theoretical concepts with practical applications through real life examples drawn from case studies, anchored in specific administrative and technological contexts;
- To provide university programs with content and structure for courses on digital records management and preservation.

1.5   Scope

Digital Records Pathways: Topics in Digital Preservation consists of the following modules:

Module 1:   Introduction – A Framework for Digital Preservation
Module 2:   Developing Policy and Procedures for Digital Preservation
Module 3:   Organizational Culture and its Effects on Records Management Selection and Appraisal of Digital Records
Module 4:   An Overview of Metadata
Module 5:   From *Ad Hoc* to Governed – Appraisal Strategies for Gaining Control of Digital Records in Network Drives
Module 6:   E-mail Management and Preservation
Module 7:   Management and Preservation of Records in Web Environments
Module 8:   Cloud Computing Primer

Each module consists of some or all of the following components as appropriate:

- **Overview** of the topic and scope of the module;
- **Learning objectives** and expected level of knowledge upon completion;
- **Methodology** or the procedures to follow in order to apply the module;
- **Templates (where appropriate)** to facilitate the implementation of the module;
- **Case Study(ies)/Scenarios (where appropriate)** that provide real-world examples of module topic
- **Exercises** covering key learning points;
- **Review questions** to enhance comprehension and understanding of the topic;
- Additional **Resources** for the topic, including **readings, standards** and other **templates** for reference

Overview of the set			
1. A Framework for Digital Preservation 2. Developing Policy and Procedures for Digital Preservation			Foundational
3. Organizational Culture	4. An Overview of Metadata	5. Appraisal Strategies	General purpose
6. E-mail	7. Websites	8. Cloud Computing	Specific purpose
International Terminology Database			Foundational

1.6 International Terminology Database

The terminology used in the modules reflects common usage in archival and records management communities of practice. To ensure common understanding, and minimize potential confusion that may arise from regional or jurisdictional practice, all modules are supported by the International Terminology Database, available at http://www.web-denizen.com/. As well, certain specific terms are included in short glossaries in each module.

Module 7: Management and Preservation of Records in Web Environments

2   Introduction

The ubiquitous nature of the web has resulted in the proliferation of organizational websites, many of which contain information and records valuable to organizations. Websites have moved beyond acting as repositories of information to becoming sites where records are created, transactions take place and dissemination occurs. This raises records management and preservation issues unique to these environments. Because of the nature of web environments ensuring the authenticity of records created and residing on websites can be difficult.

Many large organizations have been instrumental in developing methods for Website capture and preservation tools and methodologies.[1] Many of these tools and techniques are easily adaptable to the needs of small and medium sized archival organizations or programs. The Internet Archive has been developing open source solutions for remote harvesting operations that do not require a monetary output, but do require fairly extensive technological knowledge. The National Archives of the United Kingdom have conducted research into best storage medium, a simple guide to Archiving Websites, as well as researching optimum file formats for data creation. The National Archives of Australia has produced research on metadata requirements that are key to effectively managing all digital records, including records of Web-based activity. They have also researched solutions for recording evidence of Web-based records on frequently changing Websites when infrequent crawls are in place.

2.1   Aims and Objectives

The objective of this module is to introduce key issues involved in the management and preservation of websites, including identifying records in web environments. It will aid users in recognizing the management and preservation needs of these records, identify the issues involved in managing records in a web environment and introduce strategies for website preservation.

2.2   Learning Outcomes

Upon completion of this module, you will be able to:

- Identify records in web environments;
- Understand the key issues involved in the management and preservation of records in web environments;
- Understand the different strategies for preservation of Websites;

---

[1] Among the most useful large organizations currently preserving Websites were the Library of Congress, the Internet Archive, the National Archives UK, and the National Archives of Australia.

- Know where to locate additional information and resources that will facilitated how you manage and preserve records in web environments

2.3   Terminology

This section identifies and defines key concepts/constructs that are used throughout this module.

See the ICA International Terminology Database at www.web-denizen.com for more terminology relevant to this module.

Backward compatibility: Compatible with earlier models or versions of the same product. A new version of a program is said to be backward compatible if it can use files and data created with an older version of the same program. A computer is said to be backward compatible if it can run the same software as the previous model of the computer.

Client side collection methods: The source from which the Website is collected for preservation. Client-side collection is collected via the Web browser.

Direct transfer: Acquiring a copy of the data directly from the original source. Requires access to host server. Involves copying data from the server and transferring them to the collecting institution.[2]

Remote harvesting: The most common Web archiving technique uses Web crawlers to automate the process of collecting Web pages. Web crawlers typically view Web pages in the same manner that users with a browser see the Web, and therefore provide a comparatively simple method of remotely harvesting Web content.

Server-side collection methods: The source from which the Website is collected for preservation. Server-side collection is collected via the Web server.[3]

Web crawler: A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner.

Website mirroring: A mirror is an exact copy of a data set. It essentially works as a digital "print out" of the Website. The process of Website mirroring produces a copy of the original Website, but does not capture associated metadata.

---

[2] Adrian Brown, *Archiving Websites* (London: Facet Publishing, 2006).
[3] Ibid.

3 Website Preservation Strategy

There is no single definitive solution to be applied to Website preservation. Strategies will depend upon a variety of factors including the presence (or absence) of records on the site, content ownership, technological capabilities, costs and storage abilities. Therefore, there are several action plans that could be devised for the long-term preservation of an institutional Website. The action plans range from extremely technical solutions that are highly effective and address the dynamism of certain back-end database driven Websites to simple relatively inexpensive solutions that preserve a snapshot of the Website in time.

A number of tools are available that facilitate Website archiving. The tool chosen will depend greatly on how much information the preserving organization wishes to preserve, the technical abilities of staff, and a thorough risk assessment. An approach that is based on good management practices and begun as early as possible in the lifecycle of the digital resources will be effective at least for the short to medium term.

There are many considerations for an organization about to embark on a Website preservation program. Factors include: technological abilities; rights management; training; resource description, documentation and access; choice of file formats; validation checks; disaster recovery planning; storage medium; standards; and Website capture method.

3.1 Technological Capabilities

Some strategies require an intensive knowledge of the technological environment in order for them to be implemented, while others require a minimal amount of knowledge to implement and succeed. Websites that comprise static documents and incorporate little or no interactivity are relatively simple to deal with. However, sites that incorporate high levels of interactivity and comprise dynamically generated pages are very complex and prove more difficult to archive effectively.

It is important the person(s) responsible for the preservation of digital materials have some understanding of what is involved. The individual responsible must be knowledgeable enough to have an informed exchange with those involved in the preservation strategy as well as being able to set forth realistic requirements to a third party. Additionally, within the organization, the following technological features should be considered when evaluating the appropriate web preservation strategies. These include, but are not limited to:

- Type of Website (e.g. static, dynamic);
- Server space allocations and availability;
- Back up capabilities;
- Computer system in use.

3.2   Policy / Recordkeeping Requirements

As your organization's records professional, you should already be aware of the records management principles and practices currently practiced at your organization. This module, and the others that accompany it, offer an opportune time for you to review the effectiveness of these practices and the currency of any accompanying documentation. For example, you may want to revisit retention and disposition requirements for some records series; update policies and procedures related to records management functions; or reconsider the strengths and weaknesses of any educational tools, presentations, or documents you use to instruct employees about the importance of records management or records management practices.

Policies, procedures and criteria for a Website preservation program are critical in the emerging digital environment. They ensure that the aims and objectives of the institution are carefully considered and reviewed; that collections development supports the institutional mission and priorities; and ensure accountability to the funding agencies and the wider community. Elements to consider including in a policy are: a policy statement, the goals and objectives of the policy, related documents and or legislation, scope of the policy, persons responsible for policy implementation, scope of collections, coverage, an outline of digital resource types accepted, rejection criteria, evaluation criteria, viability, and collection levels. These may be broken up into more than one policy.

See "Digital Records Pathways: Module 2: Developing Policy and Procedures for Digital Preservation" for information on policy development.

Depending on the policy hierarchy of your organization, new documentation about managing records in Web environments and Website preservation may be a standalone document or it may be part of a broader policy. In addition to the elements outlined in the Policy Module, Website management and preservation documentation should include additional elements or sections. Overall, how you structure your policy will be based on your organization's contextual information.

3.2.1   Recordkeeping

All data associated with the preservation of Websites should be included in retention schedules that govern the institution's records. Web pages should be subject to the same records management controls as other electronic records, since they provide evidence of the online activities of the organization. In addition to improved records management, the organization would benefit in terms of costs associated with storage if effective disposition schedules were in place. To ensure long-tem accessibility of data it is essential that storage media is refreshed on a regular basis. If the organization stores each iteration of the Website indefinitely then the costs associated with refreshing media will soar over time as the data collected grows.

3.3   Metadata

Metadata is the key to effectively managing all records, including records of Web-based activity. The Australian *Guidelines for Archiving Web Resources*[4] describes suggested metadata requirements for different scenarios.

For individual records on Websites and for other records of Web-based activity, this means using metadata to describe:

- Date and time of creation and registration of the record into a recordkeeping system;
- Organizational context;
- Original data format;
- The use made of the record over time, including its placement on a Website;
- Mandates governing the creation, retention and disposal of the records; and
- Management history of the record following creation – including sentencing, preservation and disposal.

For copies or snapshots of entire collections of Web resources, metadata should include:

Date and time of capture;

- Links to the universal resource indicator (URI) including information about version and date of link to specified URI;[5]
- Technical details about the Website design;
- Details about the software used to create the Web resources;
- Details about the applications (including search engines) that supplement the Web resources; and
- Details about the client software needed for viewing the Web resources[6]

It is recommended that a metadata audit be performed when embarking on a Website preservation program. This will ensure that captured resources have sufficient metadata attached to effectively preserve the accuracy, authenticity, reliability, accessibility and disposition of the resources and allow access and preservation activities to occur.

> *See "Digital Records Pathways: Module 4: Overview of Metadata" for information about metadata.*

---

[4] National Archives of Australia, "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government," (March 2001). Available at: www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf

[5] The Australian *Guidelines for Archiving Web Resources* distinguish between a URI, URL, and URN thus: Universal resource indicator (URI) a general purpose namespace mechanism; Universal resource locator (URL) an instance of URI that is the address of some resource, accessible by means of a protocol such as HTTP; Universal resource name (URN) an instance of URI that, unlike a fragile URL, is guaranteed to remain available (Jon Udell, *Practical Internet Groupware* (Sebastapol, CA: O'Reilly, 1999), 471.)

[6] "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government," from the National Archives of Australia, p. 17-18.

3.4 Rights Management / Intellectual Property Rights

Issues surrounding intellectual property rights, such as copyright concerns and moral rights have a substantial impact on any digital preservation process. "The intellectual property rights issues in digital materials are … more complex and significant than for traditional media and if not addressed can impede or even prevent preservation activities."[7] It may not only be content, but any associated software, which may be subject to intellectual property rights. "Simply copying (refreshing) digital materials onto another medium, encapsulating content and software for emulation, or migrating content to new hardware and software, all involve activities that can infringe intellectual property rights unless statutory exemptions exist or specific permissions have been obtained from rights holders."[8] Due to the nature of digital materials, strategies for continuing preservation and access may necessitate the migration of the materials into new forms or an emulation of the original operating environment. Such activities may require permissions from rights holders to legally undertake such strategies.

A specific area that could potentially become problematic is in the area of Copyright Law. According to the Canadian Heritage Information Network (CHIN), "Copyright protects the expression of ideas that are fixed in any form of media."[9] This includes various Website components, such as images appearing on a given site and the underlying software programming code:

Copyright protects the majority of creations including, literary, dramatic, musical and artistic works, sound recordings and audio-visual works. Photographs are considered artistic works. Computer software programs including underlying code have been identified as literary works and they are therefore also protected by copyright. Except where works are created in the course of employment in the course of an employee's duties or where copyright has been assigned in writing to someone else, the author of the work is the copyright holder.[10]

Copyright holders should be established and permissions granted before embarking on a Website preservation program.

3.5 Staff Development and Training

Carefully designed staff training and continuous professional development can play a key role in successfully managing any digital preservation program. All those responsible for digital preservations must have a degree of knowledge on the topic. Staff development and training can range from keeping up to date with the literature and new developments

---

[7] Maggie Jones and Neil Beagrie, *Preservation Management of Digital Materials. A Handbook* (London, UK: The British Library, 2001), 32.

[8] Ibid.

[9] Rina Elster Pantalony, *Protecting your Interests: a legal guide to negotiating Website development and virtual Exhibition Agreements* (Ottawa, Canada: Minister of Public Works and Governments Services Canada, 1999), 13.

[10] Ibid.

to participating in workshops and training modules put on by various institutions and organizations such as archival associations and educational institutions.[11]

3.6 Resource Description, Documentation and Access

Some form of classification description is essential in order to manage any archival collection and make it accessible to users; this is no different for digital collections. Major cataloguing standards, such as MARC 21 and ISAD(G), have been successfully applied to the description of archived Websites. Cataloguing and classifying archived materials allows users access to them.

Resources should be supplied with appropriate and sufficient documentation to satisfy the requirements for informed use by members of the research community. The documentation should relate to both the content and the technical format of the resource. Documentation should also provide information about the context in which resources were created and maintained before preservation, and about the relationships between the digital resource and other information sources.

3.7 Disaster Recovery Planning

The development of a disaster recovery plan that is based on sound principles, has buy-in from management and can be activated by trained staff will greatly reduce the severity of the impact of disasters. The plan will need to address the restoration of both the content of the archive, and the technical and operational infrastructure required to support it. Elements to be included in a plan should be:

- Ensure staff are trained in counter disaster procedures;
- Create archives copies of data resources each time a collection of materials takes place; Store archived copies on multiple media;
- Store archived copies on and off site;
- Complete documentation of the hardware and software infrastructure as well as operating procedures and manuals;
- Copies of all software required to operate the systems.

It is also important to test the plan to discover any issues that may have been overlooked before the event of a disaster occurs. This is also helpful to staff to allow them to become familiar with the procedures before hand. As with most policies, it is recommended that the disaster recovery plan be revisited as systems and circumstances change.

---

[11] The Society of American Archivists is one institution that organizes many workshops and Web seminars. For a calendar of current opportunities see http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/ConferenceList.html?Action=GetEvents.

3.8   Validation Checks

Once the Website has been captured and transferred to the institution's archival environment, checks must be conducted to ensure that all the parts of the Website captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all the hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process by testing to see that all links are working,[12] checking that the files can be read, checking files for completeness and accuracy and checking functionality within the files. Checks should be carried out whenever a Website archive has taken place to ensure the content and structures of the deposited data resources are intact.

3.9   File Formats

With any Website preservation program (like any digital preservation program) it is recommended that accepted file formats are defined before embarking on any collection strategy. The adoption of a single file format ensures that sustainability costs are minimized when a file format of choice is built into the records creation process.

"It has become common practice for digital records repositories, including archives, to accept certain digital file formats for long-term preservation while rejecting others"[13]. Surveys of institutions regarding file format specifications show that there are a plethora of definitions, acceptable/unacceptable formats, and preservation initiatives for file formats.[14] The PREMIS *Data Dictionary for Preservation Metadata* gives the most useful definition: "a specific, pre-established structure for the organization of a digital file or bit stream." "This pre-established structure includes how the data are encoded, which is the way in which the bits are interpreted to produce text, images and sound."[15] This is important to understand as it highlights why it is essential to specify acceptable file formats to a specific repository. "Some types of encoding are synonymous with specific file formats; for example, MP3 encoding is used to encode the MP3 File format."[16] This is simple enough to understand, but it gets increasingly complicated. Take plain text files for example, "many formats can have different encodings: even a "plain text" file can be encoded as ASCII, EBCDIC or Unicode, all of which have a number of variants."[17] The plain text file has three different types of encoding, so obviously image and music files

---

[12] See, for example: Link Checker Pro: http://www.link-checker-pro.com/; Site Audit: http://www.blossom.com/site_audit.html; Cyber Spyder Link Test: http://www.cyberspyder.com/cslnkts1.html; Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

[13] Evelyn Peters McLellan, "General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation," InterPARES 2 Project (March 2007), 1. Available at http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf.

[14] Ibid.

[15] Ibid, 2.

[16] Ibid.

[17] Ibid.

are much more complicated. "Encoding can be problematic in audio and video file formats because the optimal encoding for storage and transmission often involves compression (removing bits from the digital files to reduce their size), which can often hinder preservation efforts."[18] Further difficulties to the file format debate: "the encoding issue is further complicated by the fact that TIFF, WAVE, AVI and other common image and audiovisual formats are not file "formats" per se, but rather file "wrapper formats" (also called container formats), which are designed to combine multiple bit streams into a single file."[19] Encoding, compression and bit stream combinations all complicate how file formats are preserved over the long-term. These are also reasons why many institutions call for open formats that are well documented to ensure that sufficient documentation is available to give the collecting institution a chance of preserving digital records for the long-term.

Adrian Brown of the National Archives of the United Kingdom has identified criteria to consider when selecting file formats for data creation. The criteria include:

- Ubiquity
- Support
- Disclosure
- Documentation quality
- Stability
- Ease of identification
- Intellectual property rights
- Metadata support
- Complexity
- Interoperability
- Viability
- Re-usability

Although the research does not recommend actual file types, these criteria are important to bear in mind when selecting file formats.[20]

It is important that the creating and preserving organization(s) develops policy that states the types of file formats that are acceptable to preserve. By restricting the range of file formats that an institution agrees to receive and manage, the organization can be assured that the file formats it collects adhere to the criteria stated above and that they adhere to current standards. If "good" file formats are collected, the difficulties in preserving them will be minimized as well as costs reduced.

---

[18] Ibid
[19] Ibid.
[20] Adrian Brown, "Selecting File Formats." Available at
http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

"Many formats are proprietary, that is, they are the property of an owner who, for commercial reasons, is not willing to provide access to documentation about them, and who may require a fee to be paid for their use."[21] This is a reason why most experts recommend file formats that adhere to open standards. This is also a reason why many file format registries have been developed. The registries exist to provide reliable and detailed information about file formats. Examples of file format registries include: PRONOM[22] and the Global Digital Format Registry.[23] In April 2009 the Global Digital Format Registry initiative joined forces with the UK National Archives' PRONOM registry initiative under a new name - the Unified Digital Formats Registry (UDFR). The UDFR will support the requirements and use cases compiled for GDFR and will be seeded with PRONOM's software and formats database.[24]

The collecting organization can help promote sound records creation by publicizing those file formats that are most likely to be sustainable over a period of time and by encouraging records creation using these particular formats. Another alternative is for the collecting institution to convert all digital materials preserved to the file format of choice once the material is in the archives.

3.10   Storage Medium[25]

Whichever capturing method is used, the Website needs to be preserved and stored on a relatively stable electronic digital medium. Currently, no electronic digital medium can be considered archival due to concerns regarding the relatively short and/or unproven life spans of such media and to concerns regarding technological obsolescence resulting from rapid changes in the technological environment. Storage hardware is being continually developed. Current "state of the art" medium may be obsolete in fie years time and simply impossible to maintain in 20 years time. Electronic media are not as permanent as is often thought. Manufacturers may claim satisfyingly long lifetimes for their media26 but practical experience suggests that a realistic figure for the life of a magnetic tape may be 15 years, and for a CD 20 years, all depending on original quality, storage, handling, and usage. And even if the media lifetime is longer, the hardware to read it may not be available. For many media, a small imperfection that appears after some time may make

---

[21] Ross Harvey, *Preserving Digital Materials* (Munich: K. G. Saur, 2005), 141.
[22] PRONOM is a file format registry established by the National Archives (UK) to provide and manage information about file formats and software applications used. The PRONOM Website can be found at: www.nationalarchives.gov.uk/pronom.
[23] The Global Digital Format Registry was also developed to support digital preservation. http://www.gdfr.info/.
[24] The Unified Digital Formats Registry is available at: http://www.udfr.org/.
[25] In this report we present basic storage medium for storing electronic media. It is possible to create a repository for digital materials. zIf you require more information take a look at the ISO Standard: ISO 14721: 2003, more commonly known as the Open Archival Information Systems (OAIS) reference model and OCLC and NARA. "Trustworthy Repositories Audit & Certification: Criteria and Checklist" Version 1.0, 2007. Available at: http://www.crl.edu/PDF/trac.pdf.
[26] 1995 Kodak research on their writeable CDs, reported at http://www.cd-info.com/CDIC/Technology/CDR/Media/Kodak.html, quoted a lifetime of 217 years under specified conditions.

the whole medium unusable.27 Therefore, whichever medium is chosen for storage will need to be periodically checked and/or refreshed to counteract data loss.[28]

A variety of factors affect the longevity of electronic media, including storage conditions, quality of the products used, and the composition of the products due to the availability of better materials over time. Therefore, it is difficult to predict longevity. The Canadian Conservation Institute has put together a table that provides estimates of predicted longevity for various media storage types.

[27] Jim Liden Sean Martin, Richard Masters and Roderic Parker, "The large-scale archival storage of digital Objects," DPC Technology Watch Series Report 04-03, February 2005.
[28] See The National Archives of the UK's Digital Preservation Guidance Note: 2, "Selecting Storage Media for Digital Preservation," by Adrian Brown, Head of Digital Preservation Research, August 2008. Available at: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf.

Predicted longevity of electronic media29

Media type	Predicted longevity
Magnetic disks	
Hard disks	2–5 years
Floppy diskettes	5–15 years
Magnetic tapes	
Digital	5–10 years
Analog	10–30 years
Optical discs	
CD-RW, DVD-RW, DVD+RW	5–10 years
CD-R (cyanine and azo dyes)	5–10 years
Audio CD, DVD movie	10–50 years
CD-R (phthalocyanine dye, silver metal layer)	10–50 years
DVD-R, DVD+R	10–50 years
CD-R (phthalocyanine dye, gold metal layer)	>100 years
Other optical discs	
MO, WORM, etc.	10–25 years?
Flash media	unknown

It is therefore recommended that the archived Website be stored in several environments—for example, on a hard drive and on DVD-R—and stored in the archives to counteract these storage concerns and help assure long-term access to the stored data.

In determining what type of storage media to store digital materials a number of factors need to be considered. These factors include longevity, capacity, viability, obsolescence, cost and sustainability, documented by Adrian Brown at the National Archives of the United Kingdom.[30] Brown displays a scorecard comparing common media types:

---

[29] Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives.* Available at: http://www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx.
[30] The National Archives, "Digital Preservation Guidance Note 2: Selecting Storage Media for Long-Term Preservation," August 2008. Available at: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf.

Media	CD-R	DVD-R	Hard disk	Flash Memory Stick and Card	Linear Tape Open (LTO)
Longevity	3	3	2	1	3
Capacity	1	3	3	2	3
Viability	2	2	2	1	3
Obsolescence	1	2	2	2	2
Cost	3	3	1	3	3
Susceptibility	1	1	3	1	3
Total	**11**	**14**	**13**	**10**	**17**

According to this chart, the top two storage solutions are Linear Tape Open and DVD-R, with a hard drive option a close third. Brown advises:

> In situations where multiple copies of data are stored on separate media, it may be advantageous to use different media types for each copy, preferably using different base technologies (for example, magnetic and optical). This reduces the overall technology dependence of the stored data. Where the same type of media is used for multiple copies, different brands or batches should be used in each case in order to minimize the risks of data loss due to problems with specific manufacturers or batches.

Joe Iraci, of the Canadian Conservation Institute, has additional comments regarding the differences of storage media. With regard to using optical storage media for storage, Iraci states: "the type of disc chosen and how it is recorded greatly impact[s] longevity." He highlights that "digital tapes have short lifetimes and need to be migrated/refreshed every 5-10 years" warns that "hard drives are not for long-term storage and data needs to be moved to a new hard drive every 2 to 5 years" and reminds us to "stick with technologies that are in widespread use and avoid new technologies" such as "Blu-Ray, Holographic Storage [and] Flash Media." Iraci also points out that "With all digital media, backups are critical in order to avoid sudden loss of information."[31]

Research such as that conducted by Adrian Brown and the Canadian Conservation Institute is invaluable when deciding what media to choose for the storage of institutional electronic records. It is clear that a variety of media should be chosen and that even with correct storage and handling the medium should be checked and refreshed regularly.

3.1   Standards

A number of standards are related to Website archiving. HTML and XML are core technologies recognized as standards in the form of W3C[32] recommendations. Two standards exist in the area of records management: ISO 15489-1/2: 2001 sets standards for records management practice, ISO 23081-1: 2006 sets standards for records management metadata.

---

[31] E-mail from Joe Iraci to Randy Preston, May 20, 2009.
[32] W3C or the World Wide Web Consortium is an international consortium where Member organizations, a full-time staff, and the public work together to develop Web standards.

ISO 14721: 2003 sets the standard for defining fundamental requirements for a digital preservation system. More commonly known as the Open Archival Information Systems (OAIS) reference model, its concepts and terminology have been widely adopted by an international audience. It forms the basis for the certification scheme for trusted digital repositories.

> *See "Digital Records Pathways: Module 1: Introduction – A Framework for Digital Preservation" for more information on the OAIS reference model.*

ISO 19005-1: 2005 or the PDF/A standard has addressed the need for open digital file formats. The standard is "a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rending the files."[33]

3.12   Maintaining Web-based Records over Time

Ensuring accessibility to Web-based materials over time raises the same accessibility issues as surround other electronic records. There are steps that can be taken to mitigate these issues including ensuring materials are carefully managed (including maintaining the trustworthiness of Web records and identifying and mitigating the management risks), planning for obsolescence, the use of widely supported standards, implementing security measures to protect against either deliberate or accidental alteration, and ensuring environmental control and monitoring. Most of these steps have been discussed in previous parts of this document, but it is prudent to stress again the importance of these issues.

Careful Management: This might include: maintaining preservation masters and storing these in a separate location; implementing the use of XHTML and avoiding the use of non-standard HTML tags, refreshing storage media regularly, spot checking data to ensure accessibility.

Planning for Obsolescence: Plan for obsolescence by ensuring that records can be copied, reformatted or migrated. Any preservation activities such as the above should be documented in the recordkeeping metadata, including any loss of functionality, content or appearance.

Use of Standards: The importance of the use of standards has been documented above.

---

[33] ISO-19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).

Security Measures as a means to Protect Data: It is important to build into the Web preservation process security measures that protect data from either deliberate or accidental alteration. Measures can be as simple as keeping the preserved data in a secure environment that has controlled access to allow only authorized persons access to the data and providing read-only access to the preserved data.

Environmental Control and Monitoring: Best practice dictates that stored media should be kept in optimal temperature and humidity levels, media should be protected against magnetic fields, the use of air filtration units to protect against air pollutants, prohibiting the consumption of food in the storage area, and planning for disasters.

> *See NARA Guidance on Managing Web Records, available at: http://www.archives.gov/records-mgmt/policy/managing-web-records-index.html*

3.13   Website Capture Methods/Tools

There are two types of Websites – static and dynamic. A static Website is composed of a series of pre-existing Web pages, all of which are linked to from at least one other page. A dynamic Website generates Web pages on-the-fly from smaller elements of content. Such content can be housed in a database, drawn from external sources and inserted into a Web page, or generated by scripts that respond differently depending on such factors as the date or time the Web page is accessed.

Currently, there are three options available for capturing Websites. The methods for capture vary depending on how much information the collecting institution wishes to preserve. Information includes functionality, metadata and the degree of authenticity, reliability and accuracy the collecting institution wishes to preserve. The three options are: direct transfer, remote harvesting and Website mirroring.

3.13.1 Direct Transfer

The only way to fully recreate a Website in a preservation environment is through Direct Transfer of data. Direct transfer works by acquiring a copy of the data directly from the original source. This requires direct access to the host Web server. Direct transfer then involves copying the selected files from the server and transferring them to the collecting institution. To guarantee continued functionality minor adjustments may need to be made to the preserved site.[34] To ensure that the preserved Website is as authentic as possible, a recreation of the technical environment in which the Website resides will need to be implemented within the archival setting. This means that the database or content

---

[34] For example: The hyperlinks within the archived site may need to be adjusted from absolute links to relative links; and the appropriate search engine (the one used in the original environment) must be installed in the new environment to ensure that search functionality is preserved. For a more comprehensive explanation see: Adrian Brown, *Archiving Websites*.

management system will need to be installed in the archival environment, together with the necessary Web server and search engine software. Direct transfer is the only method that takes into consideration the dynamic nature of a Website and is the only way to preserve all possible forms of dynamically generated data. However, the implementation and support of such a method will require staff with appropriate technical skills be available to install and maintain the system.

3.13.2 Remote Harvesting

The remote harvesting solutions offers three alternatives: a straight forward automated crawl of the Website, a "snapshot" crawl with additional logs kept by the archivist to back up the data mined in the snapshot, and outsourcing the process to a third party. Remote harvesting collection methods as alternatives must be understood with the caveat that such data collection methods do not capture the entirety of all Web page possibilities that could be generated by a user request, if the Website identified for capture is a dynamic site with an underlying back-end database used to house information generated on the fly. Also, using this method may result in the presence of broken links within the copied data environment as pages may contain links to content that needs to be generated on the fly to appear for the user. Other data loss that could occur may be loss of graphics and the template design.

A snapshot of a Website usually involves creating a full and accurate copy of an organization's Website at a particular point in time. A snapshot should include all aspects of the Website to ensure that a fully functional site can be recreated. The snapshot should include scripts, programs, plug-ins, and browser software—components that make the snapshot fully functional.

A standard Web crawl could be conducted using an open source Website harvesting software, such as GNU Wget free utility or Heritrix.

The advantages of an open source crawler for Website archiving are that it is non-proprietary and therefore no financial penalties would be incurred. An automated Web crawl could collect data as frequently as the institution desires; initially the crawler could be set to crawl the entire site, and subsequent crawls could collect data from pages that have only been updated since the previous crawl.

To preserve an impression of the Website at a given moment in time, the institution need only crawl a Website once or twice a year. This frequency, however would obviously not capture every change made to a Website, and may miss some of the documented activity that is present. The Web crawler would be implemented to perform infrequent crawls of the Website. Copies or "snapshots" of the Website as a whole are taken (ensuring that the functionality of internal links are not destroyed and are maintained). In the meantime, to ensure that the necessary evidence is captured a log of changes that determines when and how documents or Web pages are removed, replaced or updated, is kept. If, for the purposes of accountability and site maintainability, it is important that records of Website

content and changes are made and kept, then this is a viable, inexpensive option.[35] Once again, metadata is the key to effectively managing all records, including records of Web-based activity. (See previous Metadata heading).

There is the option of outsourcing the capture and storage of Websites to fee based companies. Services such as Web Archiving Service (WAS) developed by the California Digital Library and the Archive-It project run by the Internet Archive provide web capture and storage services to organizations that wish to preserve their Websites. It is important to note with this option that data stored by companies hosting these services can be spread across the globe and subject to a variety of jurisdictional laws and regulations. It is important before embarking on such a project to be aware of your organization's legislative and regulatory framework around data protection, privacy and access to information.

3.13.3 Website Mirroring

An option that copies the Website, but will not capture associated metadata needed to effectively preserve the digital content of the Website, is Website mirroring. A mirror is an exact copy of a data set. It essentially works as a digital "print out" of the Website. Mirroring of sites occur for a variety of reasons, one of them being to preserve a Website or Web page.

Mirroring, as stated above, does not capture metadata associated with each Web page file. It is a good option if all the Archives wishes to preserve is evidence of the Website in question. This solution should be understood with the proviso that as there is no metadata capture during the process of mirroring the Website, there is nothing in place to address evidence of actual records that may appear on the site. Therefore, it is not recommend if the collecting archives wishes to preserve evidence of records appearing on the Website.

3.13.4 Web Capture Tools

The open source crawler HTTrack has been utilized effectively in other archival institutions.[36] HTTrack is a free and easy-to-use offline browser utility. It allows a user to download a Website from the Internet to a local directory, building recursively all directories, copying HTML, images and other files from the server to the local directory. HTTrack arranges the original site's relative link-structure. It allows users to simply open a page of the "mirrored" Website in their browser and to browse the site from link to link, as if viewing it online.[37] Archivists seeking to preserve Web content in the Microsoft/Windows environment have used this harvester successfully.

---

[35] The Web crawl with a log option was researched using "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government" from the National Archives of Australia. It is a government recordkeeping document published in March 2001 and can be downloaded from http://www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf (last accessed April 28, 2009).
[36] E-mail to the Management & Preservation of Electronic Records Listserv, April 3, 2009, from the Electronic Records Archivist at Kentucky Department for Libraries and Archives.
[37] See the HTTrack Website for more information: http://www.httrack.com/.

See "Practical E-Records" for a review of HTTrack, GNU Wget free utility, Heritirix and Web Archiving Service. Available at: http://e-records.chrisprom.com/?tag=website-harvesting. The was reviewed based on the following criteria: Installation/Configuration/Supported Platforms; Functionality/Reliability; Usability; Scalability; Documentation; Interoperability/Metadata support; Flexibility/Customizability; License/Support/Sustainability/Community.

Additionally, the Adobe Web Capture tool converts Web pages to PDF files to create PDF versions of the Web page. It is simple to use and therefore easily teachable to staff. It is possible to capture an entire site using Web Capture. Not only do all the links continue to work in the PDF, they also link to local content within the PDF, where applicable, so that you can truly browse the site offline. Web Capture can be invoked through the Acrobat toolbar in Internet Explorer on Windows and through the Adobe Acrobat 9 application on Windows and Mac platforms.

The tool is easy to use, captures various levels of links within a site, has a date and time stamp for captured web pages, and backwards compatibility is assured by Adobe. There is, however, no metadata captured, it reproduces a flat PDF document, which means that it is not possible to remove a portion of page to print, for example a picture, so it becomes necessary to print the whole page, the entire Website is captured each time, and the tool converts the captured Website to PDF rather than to PDF/A.

Adobe released the Web capture tool in 2008, and is an extremely simple solution to implement and use. Adobe has a good reputation and a history of support for the client. Adobe tries to ensure that each new product release is backward compatible to several previous versions.

One further tool for Website archiving (for the archiving of Websites built with a back-end database to house the information that is generated on-the-fly to the user) is database archiving. The technique is in its infancy, but it is worth describing in some detail, as it is a tool that can be utilized to mitigate problems associated with archiving dynamic Websites using static Website methods.

Brown describes the process of archiving database driven Websites as having three stages: First the repository defines a standard data model and format for archived databases; then each source database is converted to the standard format; and, finally a standard access interface is provided to the archived databases.[38]

The Swiss Federal Archives have developed an XML based format that permits long-term preservation of relational database content. The format has a long history of development dating back to the early 1990s. In May 2008 it was accepted as the official format of the European PLANETS project for archiving relational databases. The format is known as SIARD or the Software Independent Archiving of Relational Databases. It preserves data content and metadata as well as the relations in a format that conforms to

---

[38] Brown, *Archiving Websites*, 59.

ISO standards. A briefing paper published in October, 2008 by digital preservation Europe, "Database Preservation: The International Challenge and the Swiss Solution" describes the SIARD process:[39]

A SIARD archive is a structured non-compressed ZIP container (ZIP-64 standard), permitting practically any file size. It contains two folders: "header" and "content." The header folder stores the database context, the metadata. A single file, *metadata.xml*, assures that we can understand the technical as well as the contextual background of the database. In technical terms SIARD registers on the upmost level (the database) the identifier, the format version, the message digest code of the archiving pc terminal (verifying primary data integrity) etc. On the schema level SIARD stores lists of tables, views and routines. On the table level, SIARD records the constraints and triggers. And as we go deeper into the column level SIARD also specifies the SQL type in use, LOBs (Large Objects) names, and most important of all: foreign keys and candidate keys with referential data – i.e. the relations. At the same time SIARD contextualizes the data. On the database level it lets us register or add (with the SIARD Suite) information on the archive provenance, description, user etc. In lower levels it lets us keep details of the tables and columns names and content. This descriptive information renders the database comprehensible for future users in both contextual and technical terms.

The second folder, content, stores the primary data. The data is archived according to the database structure. For each schema SIARD automatically generates a folder (schema 1, schema 2, etc.), containing the corresponding table series as subfolders (table 1, table 2, etc.). Data itself is stored in XML files (e.g. table1.xml). This schema definition reflects the table's SQL schema metadata. And it specifies that the table is stored as a chain of lines encompassing a sequence of column entries with different XML types. BLOBs and CLOBs (Binary or Character Large Objects containing all sorts of information) are also archived. They are stored in automatically generated folders (e.g. lob1, lob2, etc.) either as TXT or BIN files (record1.text, or record1.bin, etc.).[40]

SIARD is also an open format, which would mean that the collecting organization could in fact archive the database without the possible additional costs of obtaining a license to the proprietary content management system required if the Direct Transfer method of capture is employed.

---

[39] According to the briefing paper published in October, 2008 by Digital Preservation Europe, "Database Preservation: The International Challenge and the Swiss Solution".
(http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf), "The use of widely accepted ISO standards ensures to a large extent that stored data could be accessed in the future. Based upon this assumption SIARD records both primary data and metadata automatically in ISO norm formats: SQL1999 UNICODE and most important of them all: XML 1.0. To ensure standardization SIARD converts all proprietary database charters into the equivalent UNICODE character set. Furthermore, SIARD does not archive synonyms as they are not part of the standardized SQL:1999. Sticking to the standards is an iron rule."

[40] For a more comprehensive discussion of the SIARD format, please see "SIARD Format Description," available for download at the Swiss Federal Archives Website:
http://www.bar.admin.ch/themen/00532/00536/index.html?lang=en.

It is clear from the description of SIARD above, that the collecting institution will need to have input from a technologically minded individual to successfully implement the SIARD Suite.

It is uncertain at this time if the SIARD Suite is currently available for public use. At a presentation given by Jean-Marc Comment, a representative of the Swiss Federal Archives, to the 16th International Congress on Archives in July of 2008[41] it was noted that the SIARD tools will be available in the future from the Swiss Federal Archives. As of October 12, 2009 nothing appears on the Swiss Federal Archives' Website[42] regarding the SIARD tools. Due to the uncertainty of availability the SIARD Suite has not been included as a preservation option in this report. It is, however, an interesting option that may be pursued once availability is assured.

4   General Action Plan for Website Preservation

Although there is no generic solution for Website preservation plans, there are certain elements that will be universal to all programs. When the most appropriate strategy has been identified, a team comprising recordkeeping practitioners, Website administrators, communications managers, and information technology staff should be selected. The team can develop an overall action plan that includes policies and procedures that is suitable for their needs.

Following is a general action plan for Website preservation that can be adapted to many different institution's needs:

1) Identify recordkeeping requirements for Web-based activity.
2) Determine if existing system satisfies the above requirements or whether it is necessary to design and implement a new system or improve the current system.
3) Raise profile and general awareness within the organization of the general recordkeeping responsibilities of all staff.
4) Carry out risk assessment to determine level of acceptable risk posed.
5) Develop overarching Website Preservation Policy (or Digital Records Preservation Policy that includes Website Preservation)
    a) Develop Collection Policy (includes Selection policy)[43]

    b) Develop Selection Policy
        i)   Definition of context
        ii) Selection methods
        iii) Selection criteria
            (1) Appraisal of content
            (2) Extent[44]

---

[41] To view the full presentation, please visit: http://www.planets-project.eu/docs/presentations/ICA2008_Comment_SIARD.pdf.
[42] Swiss Federal Archives Website: http://www.bar.admin.ch/index.html?lang=en.
[43] Both the policy and collection lists should be reviewed periodically to add or subtract resources.

iv) Collection list
   (1) Selection of Web resources to collect
v) Define boundary definitions
   (1) Determine URL or domain name[45]
   (2) Parameters[46]
vi) Define collection method
vii) Determine timing and frequency of collection – including risk assessment methodology[47]
   (1) Influenced by life-cycle of Web resource
   (2) Rate of content change
   (3) Topicality and significance
viii)   Define storage for digital assets[48]
6) Implement Policy
7) Document procedures and processes to ensure strategies are carried out.
8) Begin Website Preservation Program.
9) Perform checks on captured and stored data.49
10) Revisit policy and appraisal objectives on a frequent basis.

---

[44] It is necessary to establish criteria for determining the extent of selected resources – i.e. whether or not external links will be collected.

[45] If a single Web resource, such as a page or a document, is to be collected in isolation, then the collection list would simply need to specify the url of that resource. If an entire Website has been selected this will usually be defined as a domain name.

[46] Parameters define the number of levels of the directory structure to be collected, and whether or not external links should be followed and if so to what depth.

[47] Cornell University has developed a methodology for assessing and mitigating risks to live Web resources: the Cornell Virtual Remote Control Project, available at: http://irisresearch.library.cornell.edu/VRC/. Abstract: The VRC risk management methodology follows a six step process that starts with the identification and the evaluation of Websites, facilitates the assessment of a site's risk level and strategy building, and initiates a subsequent response. The VRC catalogue seeks to automate this process as much as possible but allows for human control. The stability of a Website is measured at various risk levels that can be deduced from monitoring a Website over time regarding its implementation (e.g., HTML tidiness) and its hyperlink structure, as well as from metadata about the Web server (e.g., server software, response time). If a Website is at high risk it may be necessary to get in touch with the respective site owner. VRC therefore plans to establish recommendations in 'Web content preservability guidelines.' As a last resort the Web resource at risk may also be harvested and preserved to avoid its loss. Abstract from ERPANET: erpaAssessment, available at http://www.erpanet.org/assessments/show.php?id=1092153049&t=1.

[48] To ensure long-term accessibility of data, it is essential that storage media is refreshed on a regular basis; the action of refreshing storage media should be built into the overall electronic records policy seen in the action plan steps.

[49] Once the Website has been captured and transferred to the archives environment, checks must be conducted to ensure that all the parts of the Website captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process. Examples of link testing applications are: Link Checker Pro: http://www.link-checker-pro.com/ Site Audit: http://www.blossom.com/site_audit.html Cyber Spyder Link Test: http://www.cyberspyder.com/cslnkts1.html and Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

5 Case Study: Development of a website preservation plan for a student society at an academic institution

This section discusses a case study that involved the development of website preservation plan for a student association within a large academic institution. This case study offers an example of how to identify requirements and develop an organizational website preservation plan. The goal of the study was to develop strategies for exercising greater control of modifications to the Society's website and for the long-term preservation of its various iterations over time.

5.1 Background on Organization

The Acme University Student Society (AUSS)[50] is located on the campus of Acme University and is the University's student society. The society consists of 30,000 members made up of students from the main and satellite campuses of the University. The AUSS operates as an independent not for profit society. The purpose of the AUSS is to oversee services to students (e.g. tutoring, job hunting, etc.), businesses and clubs. The AUSS is the archives and records centre for the Alma Mater Society.

The AUSS archivist sought strategies for the long-term preservation of a website that is frequently changing. The archivist was interested in developing strategies for exercising greater control over modifications to the website and for the long-term preservation of its iterations over time. The result of the study is an action plan that devises strategies for control over and long-term preservation if its website.

5.2 The Challenges

The AUSS had limited resources (i.e. technological expertise, staff, time, financial resources) to develop and support a strategy for the long-term preservation of its website.

5.3 The Process of Development

The AUSS's website was identified as a body of digital material for which a preservation plan would be developed. Data were collected about the institution's context and limitations, the specific body of material, its documentary forms, technological constraints, and the functional and cultural meaning of the materials.

Identify Context

Information regarding the institution, its records and its operations was compiled through an ethnographic approach. Various interviews and observations were conducted with the Society's Archivist, Communications Manager, Website Editor and its Information Technology Manager, producing a contextual analysis, diplomatic analysis and providing information on the records created by the AUSS, and to gain a cultural perspective of those responsible for the Website.

---

[50] A fictional university and society.

Develop Website Maintenance Procedure

The development of a procedural document that outlines how the AUSS website is to maintained was undertaken. This document sets forth procedures for website maintenance and procedures that contain criteria to be followed for what can reside on the website. Such procedures would govern website content, taking into consideration restrictions that may be placed on content due, in part to the to need to adhere to administrative, juridical and legislative requirements and norms. Having such procedures in place also ensures accurate and comprehensive appraisal in the future.

Appraise Website

An appraisal of the AUSS website was undertaken to establish what should be preserved. Four key questions were asked: 1) what to capture, 2) how often to capture, 3) how much to capture, and 4) how long to preserve what is captured.

Research Best Strategy for Preserving Website Content

Research and identify the technological option(s) that meet the AUSS's appraisal objectives and its technological, financial and human resource constraints. Identify the on-going resource costs of implementing the identified technological options. The research sought to identify methods for website preservation that had been successfully implemented in other similar organizations, as well as glean knowledge built by large organizations. Additionally, methods that had not been currently implemented were investigated.

Many large organizations have been instrumental in developing methods for Website capture and preservation and these organizations were investigated for tried and tested methodologies. Among the most useful large organizations currently preserving Websites were the Library of Congress, the Internet Archive, the National Archives UK, and the National Archives of Australia. Each of these institutions was helpful in developing an understanding of what components were necessary to be included in a preservation strategy. Much of the information is easily adaptable to the needs of small and medium sized archival organizations or programs, and without this research many smaller institutions would not be able to undertake such preservation programs.

The Internet Archive has been developing open source solutions for remote harvesting operations that do not require a monetary output, but do require fairly extensive technological knowledge. The National Archives of the United Kingdom have conducted research into best storage medium, a simple guide to Archiving Websites, as well as researching optimum file formats for data creation. The National Archives of Australia has produced research on metadata requirements that are key to effectively managing all digital records, including records of Web-based activity. They have also researched solutions for recording evidence of Web-based records on frequently changing Websites when infrequent crawls are in place. The Library of Congress has also conducted research into metadata specifically for preservation (PREMIS A data dictionary and supporting XML schemas for core preservation metadata needed to support the long-term preservation of digital materials) as well as developing other metadata schema (METS

(Metadata Encoding and Transmission Standard) A metadata structure for encoding descriptive, administrative, and structural metadata that produces Encoded Archival Descriptive Finding Aids).

Develop Website Preservation Action Plan51

A website action plan that includes strategy, protocols, functional requirements, procedures and expected outcomes was developed.

---

[51] See section 3.0 for the General Action Plan for Website Preservation that can be adapted for organizational use.

6   Review Questions

1) Name three to six criteria that should be considered when selecting file formats for data creation.
2) What are the two types of websites/website content and how do they differ?
3) Name and describe each of the three website capture methods and discuss how they differ.
4) Name three large organizations that are sound resources for archiving websites and identify their expertise.
5) When should validation checks be carried out?
6) Identify three factors that affect the longevity of electronic media.
7) What are the ten main steps in the General Action Plan for Website Preservation?

7   Additional Resources

Author: Brown, Adrian
Title: Archiving Websites
Publication Date: 2006
Source/Publisher: London: Facet Publishing

This book is a comprehensive text offering practical guidance on web archiving programs. An excellent pairing of best practices and practical advice and guidance on establishing a web archiving program. From legal issues and collection methods to program management and a look at future trends, this book is a valuable resource for information and records professionals seeking to create a website archiving program.

Author: Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze and Sandra Payette
Title: Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism
Publication Date: 2002
Source: D-Lib Magazine 8(1)
URL: http://www.dlib.org/dlib/january02/kenney/01kenney.html

Author: The Library of Congress
URL: www.digitalpreservation.gov/

The Library of Congress has a variety of excellent programs and resources that offer information and examples on digital preservation requirements, research and best practice, including the National Digital Stewardship Alliance, Digital Preservation Outreach and Education and the National Digital Information Infrastructure and Preservation Program.

Author: National Archives of Australia
Title: Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government
Publication Details: March, 2001
URL: www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf

Author: Shepherd, Elizabeth and Geoffrey Yeo
Title: Managing Records: A Handbook of Principles and Practice
Publication Details: 2003
Publisher: London: Facet Publishing

This book is a comprehensive text outlining the principles of records management and its practical implementation in organizations. It is comprehensive in its coverage of records management concepts, practices and issues. Useful to both newcomers to the profession and more experienced records managers, issues covered include organizational context; classification; creation and capture of records; appraisal, retention and disposition; access and implementation. The book includes a comprehensive bibliography of records

management resources as well as lists of national and international records management standards and professional organizations for records managers.

Author: The Internet Memory Foundation
URL: http://internetmemory.org/en/

The Internet Memory Foundation is a non-profit institution that actively supports the preservation of the Internet as a media for heritage and cultural purposes.

Author: The Internet Archive
URL: http://internetmemory.org/en/

The Internet Archive developed "Archive-It" (as discussed in this module) and has archived over 150 billion web pages dating from 1996. It also hosts a blog on the development and ongoing work of the Wayback machine:
http://iawebarchiving.wordpress.com/

Author: The National Archives UK
Title: Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism
URL: http://www.nationalarchives.gov.uk/news/734.htm

The UK National Archives is a valuable source of information on website capture, particularly their program to capture UK government websites.

7.0 References

Adobe Web capture tool. Information on the product available at:
www.adobe.com/products/acrobat/

Blue Squirrel, Grab-a-Site Product Page. Available at: www.bluesquirrel.com/products/grabasite

Brown, Adrian. *Archiving Websites* (London: Facet Publishing, 2006).

Brown, Adrian (August 2008), "Digital Preservation Guidance Note 2: Selecting Storage Media for Digital Preservation." Available at:
www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Brown, Adrian, "Selecting File Formats." Available at:
www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives*. Available at: http://www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx

Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives*. Available at: http://www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx

Digital Preservation Europe (October 2008), "Database Preservation: The International Challenge and the Swiss Solution." Available at:
www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf

European Electronic Resource Preservation and Access Network (ERPANET), *Digital Preservation Policy Tool*, 2003. Available at:
http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

Greenwood, David J. and Morten Levin, "Reconstructing the Relationships between Universities and Society through Action Research," in Norman K. Denzin and Yvonna S. Lincoln, eds., *The Landscape of Qualitative Research: Theories and Issues*, 2nd ed. (Thousand Oaks: SAGE Publications, 2003), 131-166.

Harvey, Ross, *Preserving Digital Materials* (Munich, Germany: K. G. Saur, 2005).

HTTrack Website. Available at: http://www.httrack.com/

Internet Archive, Heritrix Website. Available at: http://crawler.archive.org

InterPARES 3 Project, "Case Study 09 Alma Mater Society of the University of British Columbia." Final Report.

ISO 19005-1:2005 "Document Management - Electronic document file format for long term preservation - Part 1: Use of PDF 1.4 (PDF/A-1)."

Jones, Maggie and Neil Beagrie, Preservation Management of Digital Materials A Handbook (London, UK: The British Library, 2001).

Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze and Sandra Payette (2002) "Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism," D-Lib Magazine 8(1). Available at:
http://www.dlib.org/dlib/january02/kenney/01kenney.html

Lazinger, Susan S., Digital Preservation and Metadata. History, Theory, Practice (Englewood, CO: Libraries Unlimited, 2001).

Library of Congress (United States). www.digitalpreservation.gov/

Linden, Jim, Sean Martin, Richard Masters and Roderic Parker, "The Large-Scale Archival Storage of Digital Objects," *DPC Technology Watch Series Report 04-04*, Digital Preservation Coalition (February 2005). Available at: http://www.dpconline.org/docs/dpctw04-03.pdf

McGovern, Nancy, Anne R. Kenney, Richard Entlich, William R. Kehoe and Ellie Buckley (2004), "Virtual Remote Control. Building a Preservation Risk Management Toolbox for Web Resources," *D-Lib Magazine* 10(4). Available at: http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html

National Archives of Australia, "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government," (March 2001). Available at: www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf

National Archives of Australia. (2004). "Digital recordkeeping self-assessment checklist." www.naa.gov.au/images/digitalrecordkeepingchecklist_tcm2-923.pdf

Pantalony, Rina Elster, *Protecting your Interests: a legal guide to negotiating Website development and virtual Exhibition Agreements* (Ottawa, Canada: Minister of Public Works and Governments Services Canada, 1999).

Peters McLellan, Evelyn, "General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation," InterPARES 2 Project (March 2007). Available at: http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf

Prom, Christopher J. and Ellen D. Swain (2007), "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites," *American Archivist* 70: 344-363.

Smiraglia, Richard P., *Metadata A Cataloger's Primer* (New York, NY: The Hawthorn Press, 2005).

Society of American Archivists. Conference / Workshop Calendar 2009. Available at: http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/ConferenceList.html?Action=GetEvents

Swiss Federal Archives, "SIARD Format Description." Available at: http://www.bar.admin.ch/themen/00532/00536/index.html?lang=en