The long-term preservation of Web content

Michael Day

UKOLN, University of Bath
m.day@ukoln.ac.uk

1. Introduction[*]

Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users. To date, most such initiatives have concentrated on the development of strategies and software tools for the collection of Web content and for providing current access to this content through interfaces like the Internet Archive's Wayback Machine. The International Internet Preservation Consortium (IIPC) is currently building on this legacy with the collaborative development of a set of tools that can be used for the capture of Web sites and for the navigation and searching of Web archives. The focus on collection strategies and tools is a response to what is perhaps the most significant challenge of the Web from an information management perspective. Its dynamic nature means that pages, sites and even whole domains are continually evolving or disappearing.

It is difficult to get accurate and up-to-date statistics on Web page longevity, but a range of studies hint at the ultra dynamic nature of the Web. A study by Lawrence, *et al*. (2001) cited an Alexa Internet estimate that pages disappeared on average after 75 days. Longitudinal studies of Web page persistence by Koehler (2004) found that just 33.8 per cent of a sample set of pages selected in December 1996 persisted at their original URLs by May 2003. Studies of the longevity of Web references in scientific journals show similar trends. For example, a 2003 study of Internet ci-

[*] This is the final draft of a chapter published in: Julien Masanès, (ed.), *Web archiving* (Berlin: Springer, 2006), pp. 177-199. ISBN 3-540-23338-5.

tations in three major scientific and medical journals revealed that 27 months after publication, the proportion of inactive links rose to 13 per cent (Dellavalle, *et al*., 2003). The exact proportions differ, but similar results have been noted for Web-citations in other biomedical journals (Hester, *et al*., 2004; Crichlow, Davies & Wimbush, 2004; Wren, 2004), in computer science journals and conferences (Spinellis, 2003; Selitto, 2005), and the informetrics sub-discipline of information science (Bar-Ilan & Peritz, 2004).

Web archiving initiatives deal with the ephemeral nature of the Web by harvesting selected domains or sites, thereby creating surrogates that can be used for current and future access. Current access, where this is legally possible, can be provided through initiatives own Websites, as with the National Library of Australia's PANDORA archive, or through specific interfaces like the Wayback Machine, the Nordic Web Archive's access toolkit (Brygfjeld, 2002), or the WERA (Web aRchive Access) viewer being developed by the IIPC. Longer-term access, however, will depend upon initiatives being able to preserve the Web content that has been collected, thus bringing us into the domain of digital preservation.

The remainder of this chapter will outline the challenges of digital preservation, focusing in more detail on repository systems, preservation strategies and metadata. A final section will consider some specific preservation issues as they relate to the content gathered by Web archiving initiatives.

2. The challenge of long-term digital preservation

Digital preservation can be understood as referring to the range of activities required to ensure that digital objects remain accessible for as long as they are needed. In a popular definition, Hedstrom (1998, p. 190) says that digital preservation involves "the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable." Despite the growing ubiquity of digital information, the long-term preservation of information in digital form is far from a simple task. At the heart of the problem is the rapid obsolescence of the various technologies on which digital information depends, as outlined in the influential report of a task force set up in 1994 by the Commission on Preservation and Access and the Research Libraries Group (Garrett & Waters, 1996, p. 2):

Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, "nasty, brutish and short."

As Hedstrom's definition suggests, the challenges of digital preservation are multifaceted, involving a mixture of technical and organisation issues. While most of the underlying difficulties relate to technology, successful solutions to the digital preservation problem will depend upon what Smith (2003, p. 2) describes as the "series of actions that individuals and institutions take to ensure that a given resource will be accessible for use at some unknown time."

The following sections will outline in slightly more detail the some of the reasons why digital information is difficult to preserve.

2.1 Technological challenges

One fundamental problem is the stability of the media that digital information is stored on. A comparative study of media types undertaken in the mid-1990s suggested that, given ideal storage conditions, magnetic tapes could only reliably retain data for around 20 years, while more traditional media like acid-free paper or silver-halide microform could last for centuries (Van Bogart, 1995). More recently developed storage media types may be more robust, but it is difficult to be certain about this. As Ross and Gow (1999, p. 2) have noted, "it often proves difficult to make-well informed and secure decisions about technological trends and the life expectancy of new media." In practice, dealing with media longevity means that content needs to be copied periodically to new media or new media types. This process is called 'refreshing,' and is one of the activities associated with good data management practice, like the making of regular backups.

Media deterioration, however, is not the only technical preservation issue. As we noted before, a more pressing problem - and ultimately one that is more difficult to solve - is dealing with the technological obsolescence of hardware, software and media types. As Brichford and Maher (1995) pointed out with regard to hardware obsolescence, a "twenty-year life for the plastic backing material used for computer tapes and disks is irrelevant if the tape or disk drives on which they were recorded become obsolete and unavailable after ten years." The dependencies of digital objects on hardware and software can be complex. For example, Thibodeau (2002, p. 6) views digital objects as inheriting properties from three separate classes:

Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly. A physical *object is simply an inscription of signs on some physical medium. A* logical *object is an object that is recognised and processed by software. The* conceptual *object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions.*

Strategies for dealing with the technical obsolescence problem include the periodic migration of objects to new formats and attempts to preserve or emulate technology. These will be introduced in more detail in the section on preservation strategies below.

2.2 Other challenges

In addition to these largely technical problems, there are a series of related challenges that relate to the long-term preservation of digital objects.

The first relates to the difficulties of ensuring the authenticity and integrity of objects over time. Digital information is relatively easy to manipulate, meaning that it can easily be deliberately or accidentally corrupted (Lynch, 1996). The users of digital resources need to have confidence in the authenticity of preserved objects, i.e. that they are what they claim to be and that their integrity has not been compromised. There are technical methods available for dealing with this issue at the bit-level (e.g. cryptographic techniques), but confidence in an object's authenticity will ultimately be based on the level of trust a user has in the repository responsible for maintaining it.

A second problem relates to scale, i.e. the massive (and growing) amounts of digital information now being generated, combined with a proliferation of format types. The Web is but one exemplar of this, another being the 'data deluge' now apparent in many scientific disciplines, whereby vast amounts of data are being generated by high-throughput instruments or streamed from sensors or satellites (Hey & Trefethen, 2003; Szalay & Gray, 2006). Because Web archives tend to collect multiple snapshots of Web content, they can grow very quickly indeed. For example, the largest current initiative, the Internet Archive, provides access to approximately two Petabytes of data and is growing at the rate of 20 Terabytes a month.[1] On the other hand, some national Web domain crawls can have compara-

[1] Figures are taken from: Internet Archive, Frequently Asked Questions. Retrieved May 31, 2006 from http://www.archive.org/about/faqs.php

tively modest storage requirements. For example, Hakala (2004) reported that a crawl of the Finnish Web domain in 2002 collected a total of 500 Gigabytes. A crawl of the Portuguese Web in 2003 processed 3.8 million URLs and downloaded 78 Gigabytes of data (Gomes & Silva, 2005). By contrast, the first domain harvest of the Australian Web in 2005 took six weeks and captured 185 million documents or 6.69 Terabytes of data (Koerbin, 2005).

A final set of challenges relate to the legal contexts of digital preservation. So, for example, intellectual property rights (IPR) legislation or restrictive licensing mechanisms can sometimes restrict the collecting and preservation activities of cultural heritage organisations. Indeed, López Borrull and Oppenheim (2004) have noted that recent changes in IPR law have tilted the balance of rights away from users in favour of content owners. While carefully constructed legal deposit legislation can help to solve some of these challenges, some preservation strategies depend on the adaptation (or reengineering) of application programs in ways that would not normally be permitted by software licenses. IPR issues are, however, not the only ones that are relevant. A detailed study of the legal contexts of Web archiving by Charlesworth (2003) noted significant problems with the potential liability of archives for providing access to defamatory or otherwise illegal content, or for breaches of data protection laws. As it is unlikely that all of these legal challenges are going to be solved in the short term, it is important that those responsible for Web archiving and digital preservation activities maintain a watching brief on legal developments in their respective legal jurisdictions.

3. Developing trusted digital repositories

The challenge of technical obsolescence means that traditional preservation activities focused on maintaining and conserving objects are no longer effective for supporting the long-term preservation of digital information. Instead, organisations that need to preserve digital information have to develop processes and systems that can be implemented over a long period of time, adapting to external developments as necessary. This 'active' approach to preservation is typified by the definition of digital preservation proposed by a working group sponsored by the Research Libraries Group (RLG) and OCLC Online Computer Library Center: "the managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents" (RLG/OCLC Working Group on Digital Archive Attributes, 2002, p. 3). These managed activities depend upon the existence of an organisational entity that can take responsibility for maintaining digital objects. In practice, this means developing some kind of repository or archive. The *Trusted Digital Repositories* report, produced by the working group referred to previously, defines a trusted repository as, "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future" (RLG/OCLC Working Group on Digital Archive Attributes, 2002, p. 5). Such repositories have to undertake a number of different functions. A start in defining some of these has been made by the OAIS Reference Model.

3.1 The OAIS Reference Model

The Reference Model for an Open Archival Information System (OAIS) is an attempt to provide a high-level framework for the development and comparison of digital archives or repositories (CCSDS 650.0-B-1, 2002). This standard - which has also been approved by the International Organization for Standardization as ISO 14721:2003 - was developed by the Consultative Committee for Space Data Systems (CCSDS) as part of an initiative to develop standards that would support the long-term preservation of data retrieved from satellites and other kinds of space mission. Despite these domain-specific origins, OAIS has been developed as a generic model, applicable in many other digital preservation contexts.

The OAIS model aims to provide a common framework that can be used to help understand archival challenges, especially those that relate to digital information objects. The value of the model is in providing a high-level common language that can facilitate discussion across the many different communities interested in digital preservation. The standard itself defines an OAIS as an organization of people and systems that have "accepted the responsibility to preserve information and make it available for a Designated Community" (CCSDS 650.0-B-1, 2002, p. 1-11). The Working Group on Digital Archive Attributes (2002) commented that this understanding of an archive as a system of people and systems meant that the OAIS model built a stage for a better understanding of the full requirements of digital repositories.

Before exploring the functions of an OAIS in more detail, the standard defines six mandatory responsibilities that should be discharged by an archive (CCSDS 650.0-B-1, 2002, p. 3-1).

- Negotiating with and accepting appropriate information from producers
- Obtaining sufficient control of the information provided to enable its long-term preservation

- Determining which communities constitute the 'designated community' - an OAIS concept for an identified group of potential consumers (users) who should, therefore, be able to understand the information
- Ensuring that the information to be preserved is independently understandable - i.e., without the assistance of experts - by the designated community
- Following documented policies and procedures that ensure that information can be preserved and disseminated in an authentic way
- Making the preserved information available to the designated community

Much of the rest of the standard is taken up with the detailed specification of two models that detail the functional entities needed by an OAIS and the types of information that are exchanged and managed within it. The OAIS information model will be outlined in more detail in the section on metadata below.

The functional model outlines the range of activities that would need to be undertaken by a repository, and defines in more detail those functions described within the OAIS, in order to aid the future designers of systems and to provide a set of terms and concepts for the discussion of current systems. It defines six functional entities, each of which is broken down into more detail in UML (Unified Modeling Language) diagrams.

- Ingest - accepts submissions from producers and prepares them for storage and management within the archive
- Archival storage - for the storage, maintenance and retrieval of archive content
- Data management - for managing information about the archive and its holdings
- Administration - for the overall operation of the archive system
- Preservation planning - monitoring the environment of the OAIS to ensure the long-term preservation of archive content
- Access - supporting consumers (users) in finding and retrieving archive content

3.2 Trusted digital repositories and certification

The RLG/OCLC Working Group on Digital Archive Attributes (2002) built upon the foundations laid by the OAIS Model by developing a more detailed set of requirements for trusted digital repositories. They developed a set of seven attributes of trusted repositories, the first of which is compliance with the OAIS Reference Model (p. 13):

A trusted digital repository will make sure the overall repository system conforms to the OAIS Reference Model. Effective digital archiving services will rely on a shared understanding across the necessary range of stakeholders of what is to be achieved and how it will be done.

According to the OAIS standard itself, a conforming archive would fulfil the six mandatory responsibilities and support the information model it defines (CCSDS 650.0-B-1, 2002, 1-3). However, it is careful to emphasise that, as a reference model, it does not define or require any particular method of implementation for either. Other attributes identified by the working group largely focus on organisational requirements that lie outside the scope of the OAIS model. They include the need for repositories to demonstrate a fundamental commitment to apply standards and best practice (administrative responsibility), to be able to prove their organisational viability and financial sustainability, and to have a technological infrastructure appropriate for implementing suitable preservation strategies (technological and procedural suitability).

The working group also raised the question of audit and certification, recommending the development of a framework and process to support the certification of repositories.

This was the focus of a subsequent task force supported by RLG and the National Archives and Records Administration (NARA). In late 2005, this task force published a draft audit checklist for the certification of trusted digital repositories (RLG-NARA Task Force on Digital Repository Certification, 2005), the use of which is currently being evaluated by the Center for Research Libraries (CRL) Audit and Certification of Digital Archives project (e.g., Dale, 2005). The UK Digital Curation Centre is also collaborating with RLG by conducting audits of two repositories using the checklist. The DCC team responsible for this task argues that these audits are "designed to validate not just the appropriateness of the checklist, but to provide us with an understanding of the process and costs of its use as an audit tool" (Ross & McHugh, 2005).

4. Digital preservation strategies

The OAIS Model identifies the main functions that need to be undertaken by preservation services and defines an information model for the objects held by them. However, it does not prescribe the adoption of any particular preservation strategy. This section will introduce the main range of strategies currently proposed for supporting digital preservation and comment on their appropriateness for the preservation of Web content.

The appropriateness of a given strategy depends upon the nature of the object being preserved and the reasons why it is being preserved. This means that the choice of a particular strategy, or the exact way that it is implemented, needs careful and expert consideration by repositories. In this regard, it is interesting that a number of experimental decision support tools for preservation strategies are now being developed by research projects like the digital preservation cluster of the DELOS Network of Excellence on Digital Libraries (Verdegem & Slats, 2004; Rauch & Rauber, 2004). Other research projects are investigating ways of dynamically implementing preservation strategies with the support of Semantic Web technologies for the automatic detection of format obsolescence and other kinds of incompatibility (Hunter & Choudhury, 2006).

Thibodeau (2002) has developed a spectrum of preservation strategies ranging on a continuum from the preservation of technology to the preservation of objects. The preserving technology side of this spectrum includes strategies based on maintaining original technologies or emulation. On the preserving objects side are approaches involving levels of abstraction like the persistent archives concept developed by a research group based at the San Diego Supercomputer Center (Moore, *et al.*, 2000). Between these two extremes are strategies based on the periodic transformation (or migration) of data. The following sections will introduce selected approaches in more detail.

4.1 Preserving technology

On the face of it, the simplest preservation strategy would be to keep and maintain all original application programs, operating systems and hardware platforms. Lee, *et al.* (2002) comment that advocates of this strategy argue that it is the only way of preserving the *behaviour* as well as the look and feel of a given digital object. However, this approach crumbles in the face of rapid technology obsolescence and the impossibility of maintaining hardware over long periods of time. Feeny (1999, p. 42) has argued that this strategy would quickly result in the existence of museums of "ageing and incompatible computer hardware." While the strategy may have some value where the hardware is particularly unique or historically significant, this is certainly not an approach that is appropriate for Web content, which is usually accessed via browser software rather than being dependent on any specific type of hardware.

4.2 Emulating technology

Preservation strategies based on emulating technology abandon attempts to keep obsolete hardware working and focus instead on the development of programs that enable the continued use of application programs in new environments. A basic assumption of the strategy is that digital resources are inherently software-dependent. According to Rothenberg (1999, p. 8), "digital documents exist only by virtue of software that understands how to access and display them; they come into existence only by virtue of running this software." He argues that the only reliable way of recreating a document's original functionality, look, and feel would, therefore, be "to enable the emulation of obsolete systems on future, unknown systems" (p. 17). The importance of the emulation approach is that it enables the preservation of digital objects in their original forms, which aids their authenticity.

Emulation, therefore, is based on the development of software programs (emulators) that mimic the behaviour of obsolete hardware. Technically speaking, this is far from being a trivial task, but the fact that hardware tends to be well specified at a logical level means that it is an easier task than reengineering application software for new computing environments (Digital Preservation Testbed, 2003). Once an emulation approach has been chosen, there is a need to solve the practical question of exactly how emulator programs will be run on future generations of hardware. One suggested approach is to 'rehost' emulator programs periodically onto new hardware platforms, which could be quite resource intensive. An alternative is 'chaining,' whereby emulator platforms are in time themselves successively emulated, enabling previous emulators to be run under a chain of emulators. Another approach that has been suggested involves developing all emulator programs to function on a virtual platform - an 'emulation virtual machine' - that can in turn be successively implemented on new hardware platforms (Rothenberg, 2000). The advantage of this approach is that it simplifies the amount of rehosting required, as only the virtual machine itself needs to be rewritten. A variant of the virtual machine approach is the Universal Virtual Computer (UVC) concept developed by Raymond A. Lorie of IBM (Lorie, 2002). A UVC is a simple general-purpose computer that can be implemented as a platform independent layer on current and future hardware. Rather than running original application programs, formats supported by the UVC are decoded into a Logical Data Schema that can be used with format decoders in the future to reconstruct objects on future implementations of the UVC. The National Library of the Netherlands and IBM has tested the concept through the development of a UVC demonstration tool for JPEG images (Hoeven, Diessen & Meer, 2005).

Emulation would, in principle, seem to represent an appropriate preservation approach for at least some Web content. Web browsers tend to be implemented on a wide range of different platforms, so the exact choice of hardware to emulate would to some extent be arbitrary. Because Web standards and the technologies supported by browsers change over time, it may be important to maintain multiple versions of browser software in order to ensure that pages can be rendered in an appropriate manner. It will also be important to maintain copies of browser 'plug-ins' and related technologies.

4.3 Migrating objects

The data migration preservation strategy abandons any attempt to keep obsolete hardware or application programs working, even in surrogate form. Instead, data objects are continually transferred in order to work on new generations of hardware and software. The Task Force on Archiving of Digital Information (Garrett & Waters, 1996, p. 6) provides a much-cited definition:

> *Migration is the periodic transfer of digital materials from one hardware/software configuration to another, of from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.*

Data migration is currently the most tried-and-tested preservation approach, often combined with some kind of format standardisation undertaken on ingest. Migration has been used for decades by the computer industry to ensure that current information remains accessible and usable. It also underlies the versioning behaviour of formats in many office programs. From a longer-term perspective, however, it suffers from a number of problems. Firstly, because objects are subject to almost continuous change, it is very difficult to ensure that they retain their authenticity (Tibbo, 2003, p. 22). Secondly, migration processes are not particularly efficient for large collections of heterogeneous objects, which would need constant monitoring and intervention. Rothenberg (1999, p 13) argues that migration approaches are labour-intensive, "time-consuming, expensive, error prone, and fraught with the danger of losing or corrupting information."

To help overcome some of these problems, other variants of the migration approach have been developed. For example, researchers based at the University of Leeds (Mellor, Wheatley & Sergeant, 2002) proposed a form of 'migration on demand,' whereby an object's original bit-stream would be preserved - helping to maintain its authenticity - and migrated only at the point of delivery. In this model, the focus of migration moves on to the migration tools, rather than the objects themselves.

In principle, data migration approaches could be applied to the majority of Web content. The experiences of Web archiving initiatives to date suggests that much of the surface Web - at least - is made up of a relatively limited number of formats. For example, a crawl of the Finnish Web in June 2002 found that over 96% of harvested content was in one of four formats: HTML (48%), GIF (25%), JPEG (20%) and PDF (3%).[2] While it would be possible to develop data migration strategies for all of these formats - and any others that become important - the real problem is that the user experience of the Web as a whole is not so easy to migrate. The exact role of migration in supporting the preservation of Web content remains, therefore, an open question.

4.4 Other strategies

A number of other preservation strategies exist. Some of these are based on the concept of encapsulation, the idea that preserved objects should essentially be self-describing, linking content with all of the information required for it to be deciphered and understood. As we will see, this is the heart of the idea that underlies the Information Package concept in the OAIS reference model. It also underlies initiatives like the Universal Preservation Format (Shepard, 1998) and the self-documenting encapsulation concept developed as part of the Victorian Electronic Records Strategy (Waugh, *et al*., 2000; Waugh, 2006).

Another type of approach is exemplified by the persistent-archives concept developed by the San Diego Supercomputing Centre as part of a series of research projects funded by NARA and other agencies (Moore, *et al*., 2000). This is based on the development of a complete preservation infrastructure that enables the preservation of the organisation of collection as well as the objects that make up that collection, maintained in platform-independent form. The architecture used enables any hardware or software component to be replaced with minimal effect on the rest of the system.

[2] These figures are taken from a survey undertaken as part of a feasibility study into Web archiving in the UK (Day, 2003)

4.5 Final thoughts on preservation strategies

The existence of multiple approaches reflects the reality that we do not really know yet which strategies will work best for a given object or preservation objective. They are also not mutually exclusive, meaning that risk can be spread across a number of different strategies. The key thing, whichever strategy (or combination of strategies) is chosen, is to understand that the purpose of any strategy will be to ensure that the significant properties of preserved objects can be retained. In the short to medium term, however, much more research into preservation strategies will be needed (e.g., Hedstrom, 2002; Tibbo, 2003; Ross & Hedstrom, 2005).

5. Preservation metadata

The key to the successful implementation of all preservation strategies will be the capture, creation, maintenance and application of appropriate metadata (e.g., Day, 2004; 2005). This 'preservation metadata' is understood to be all of the *various types of data* that allows the re-creation and interpretation of the structure and content of digital data over time (Ludäscher, Marciano & Moore, 2001). Understood in this way, it is clear that such metadata needs to support an extremely wide range of different functions, including discovery and access, recording the contexts and provenance of objects, to the documentation of repository actions and policies. Conceptually, therefore, preservation metadata spans the traditional division of metadata into descriptive, structural and administrative categories. Lynch (1999), for example, has noted that within digital repositories, metadata should accompany and make reference to digital objects, providing associated descriptive, structural, administrative, rights management, and other kinds of information.

The wide range of functions that preservation metadata is expected to support means that the definition (or recommendation) of standards is not a simple task. The situation is complicated further by the knowledge that different kinds of metadata will be required to support different digital preservation strategies and that the metadata standards themselves will need to evolve over time.

To date, the information model defined by the OAIS Reference Model has been extremely influential on the development of preservation metadata standards.

5.1 The OAIS information model

The OAIS information model defines two main categories of metadata that needs to be associated with the objects that need information. Firstly all Information Objects handled by an archive (content, metadata, etc) are made up of a Data Object - which for digital objects would typically be a sequence of bits - and the associated Representation Information needed to permit the full conversion of these bits into meaningful information (CCSDS 650.0-B-1, 2002, p. 4-19). The OAIS model defines this Representation Information as "the information that maps a Data Object into more meaningful concepts" (CCSDS 650.0-B-1, 2002, p. 1-13), but for digital resources it is essentially the technical information (or metadata) needed to render the bit sequences into something that can be read or used by its designated community. Typically, Representation Information might include descriptions of the formats, character sets, etc. in use, possibly including software and descriptions of hardware and software environments. In the OAIS model, this is known as Structure Information. It might also include any additional information that is required to establish the particular meaning of data content, e.g. that raw numbers should be understood as dates or as temperatures in degrees Celsius. The OAIS model refers to this as Semantic Information. The OAIS information model understands that Representation Information can be recursive, i.e. that it may itself may need Reference Information in order to be made meaningful, resulting in what the model calls as a Representation Network. While Representation Information is conceptually part of the Content Information, in practice its presence could be fulfilled by a link to centralized information held elsewhere within the OAIS or even in third party registries. A start has been made with developing registries of information about file formats (Abrams & Seaman, 2003; Darlington, 2003), but similar approaches could be used for other types of Representation Information (Giaretta, *et al*., 2005).

The model also encapsulates Content Information with additional metadata - known as Preservation Description Information (PDI) to form an entity known as an Information Package. The standard defines several types of Information Package (e.g. for submission into the archive and for dissemination), but the most significant for preservation purposes is the Archival Information Package (AIP), "defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object" (CCSDS 650.0-B-1, 2002, p. 4-33). In short, the archival information package is a way of conceptually linking the object that is the primary focus of preservation together with *all* of the additional types of information (or metadata) that are necessary to support its continued use over time.

The OAIS information model says that Preservation Description Information is "specifically focused on describing the past and present states of the Content Information, ensuring that it is uniquely identifiable, and ensuring that it has not been unknowingly altered" (CCSDS 650.0-B-1, 2002, p. 4-27). The Information Model defines four separate classes of PDI, broadly based on categories defined in the report of the Task Force on Archiving of Digital Information, namely: fixity, reference, context and provenance. The report (Garrett & Waters, 1996) noted that these four categories, together with the definition of content at different levels of abstraction, were the key features for determining information integrity in the digital environment and argued that they deserved special attention.

- *Fixity* - We have already mentioned that the users of digital resources need to have confidence that they are what they claim to be and that their integrity has not been compromised. While metadata by itself cannot solve all integrity problems, the OAIS model suggests the inclusion of Fixity Information that can support data integrity checks at the level of Content Data Objects. These might include the use of cryptographic techniques like checksums that can help protect bit-level integrity by highlighting any changes made to individual data objects.

- *Reference* - Another aspect of integrity identified by the Task Force on Archiving of Digital Information was the need for objects to be identified and located over time. Their report said that for an object "to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects" (Garrett & Waters, 1996, p. 15). This brings us to the traditional realm of descriptive metadata, e.g. that used in bibliographies, catalogues, and finding aids, but also highlights a key role for persistent identifiers. Identifiers feature highly in the OAIS model's definition of Reference Information, although the practical examples make it clear that other types of descriptive metadata could also be included. There is a separate category in the OAIS information model for descriptive metadata about information packages (Descriptive Information) that can be used to facilitate discovery and access, although it acknowledges that at least some Reference Information will often be replicated in these Package Descriptions.

- *Context* - Many resources cannot properly be interpreted without some understanding of their context. Digital objects do not often exist in isolation, but interact with other objects and their wider environment. This is especially true of Web resources. The context might, for example, be technical, e.g. recording dependencies on particular hardware or software configurations. In the OAIS information model, Context Information is defined as documenting the relationships of the Content Information to its environment (CCSDS 650.0-B-1, 2002, p. 4-28).

- *Provenance* - The OAIS model understands Provenance Information as a specific type of Context Information that documents the history of the Content Information. This might include information about its creation and provide a record of custody and preservation actions undertaken. Provenance also refers to a longstanding principle of the archives profession and embodies the concept that a key part of the integrity of an object is being able to trace its origin and chain of custody (Tibbo, 2003, p. 32).

5.2 The PREMIS Data Dictionary and other standards

The first metadata specifications specifically designed to address digital preservation requirements were developed in the late 1990s by the National Library of Australia and European research projects like Cedars (CURL Exemplars in Digital Archives) and NEDLIB (Networked European Deposit Library) (e.g., Day, 2001). Between 2000 and 2002, an international working group commissioned by OCLC and RLG built upon these (and other) proposals to produce a unified *Metadata Framework to Support the Preservation of Digital Objects* (OCLC/RLG Working Group on Preservation Metadata, 2002). This Metadata Framework was *explicitly* structured around the OAIS information model, defining various metadata elements for Content Information (including Representation Information) and PDI.

Following publication of the Metadata Framework, OCLC and RLG commissioned a further working group to investigate the issues of implementing preservation metadata in more detail. The resulting Working Group on Preservation Metadata: Implementation Strategies (PREMIS) had the twin objectives of producing a 'core' set of preservation metadata elements and evaluating alternative strategies for encoding, storing, managing and exchanging such metadata.

The working group issued its proposal for core preservation metadata elements in May 2005 with the publication of the PREMIS *Data Dictionary for Preservation Metadata* (PREMIS Working Group, 2005). While this is intended to be a translation of the earlier Metadata Framework into a set of implementable semantic units, the Data Dictionary developed its own data model and was not afraid to diverge from the OAIS model in its use of terminology. The Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process," specifically that "metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS Working Group, 2005, p. ix). The Data Dictionary itself defines elements (semantic units) for describing four of the entities identified by the PREMIS data model: objects (at different levels of aggregation), events, agents, and rights, the latter two in no real detail. The working group also limited the scope of the Data Dictionary by excluding categories of metadata deemed not directly relevant to preservation (e.g. descriptive metadata) or outside the expertise of the group (e.g. technical information about media and hardware).

The examples included in the Data Dictionary include a snapshot of a Web site. This gives an indication of the potential complexity of PREMIS, in particular when it is applied to Web content.

5.3 Web archiving and metadata

Some Web archiving initiatives already collect some simple metadata. For example, initiatives using crawler programs for domain capture of the surface Web record certain aspects of documents harvested. These might include a document's original URL, its checksum, and a record of the time the document was harvested. Hakala (2004) describes how these can be used for duplicate detection and other Web archive management processes. The IIPC is current working on the development of a Web Archiving Metadata Set that would define the richer metadata that can be automatically generated or captured by IIPC tools, e.g. metadata about harvesting parameters, Web site contexts, etc.

6. Digital preservation and the Web

Before concluding, it might be worth outlining briefly some of the reasons why the Web may prove to be a particularly difficult object to preserve.

Firstly, the Web is a deceptively complex object. In governance terms it remains what Strogatz (2004, p 255) calls an "unregulated, unruly labyrinth where anyone can post a document and link it to any page at will." The result of this is hidden complexity. For example, the surface Web alone links a wide range of document types (e.g., text, images, sound, multimedia, software) in an even wider range of formats - all of which may need to be considered separately from a preservation perspective. The Web also includes (or provides interfaces to) databases, digital libraries, metadata collections, and interactive sites like 'weblogs.' In addition, while some Web site behaviour is determined at the server side (Fitch, 2003), other aspects of functionality depend on the exact combination of browser software and 'plug-ins' available to the user. In this context, it is difficult for preservation initiatives to make decisions about the significant properties and authenticity of objects. Lyman (2002, p. 41) argues that, for authenticity, preserved documents "must both include the context and evoke the experience of the original."

A related problem is the Web's dynamic nature. Web archiving initiatives can only preserve 'snapshots' of sites or domains at the expense of their dynamism, rather like insects trapped in amber. Once snapshots of Web content are located outside the active Web, it is arguably missing one of its most characteristic properties (Tibbo, 2003, p. 16).

The problems of complexity and dynamism reflect a deeper lack of clarity on defining the exact boundaries of the Web. It is a general principle of digital preservation that it is important to understand exactly what is being preserved in order to preserve it most effectively. On detailed examination, however, the Web can be a fairly nebulous concept. For example, many 'hidden Web' sites just provide browser-friendly access to a managed database whose content often predates the Web, and will most likely survive it (see chapter 5 of this book). We may need to ask whether these particular sites fall within the scope of Web archiving initiatives as currently constituted, or whether they should be dealt with in other ways.

7. Conclusions

This chapter has attempted to introduce some of the range of managed activities that are necessary to ensure the long-term preservation of collections of Web content. It has focused in particular on the development of trusted repository systems and the adoption of appropriate digital preservation strategies, noting the key role of metadata. Digital preservation has been described as a grand challenge for the first decade of the 21st century (Tibbo, 2003). Preserving Web content for the long-term promises to be one of the most demanding parts of this challenge.

Acknowledgements

UKOLN is funded by the Joint Information Systems Committee (JISC) and the Museums, Libraries and Archives Council (MLA), as well as by project funding from various sources. UKOLN also receives support from the University of Bath, where it is based.

Parts of this paper are based on a presentation given at the seminar *Archivo de la Internet española: Webs y archivos personales*, held at the Residencia de Estudiantes, Madrid on December 12, 2005.

References

Abrams, S. L., & Seaman, D. (2003). *Towards a global digital format registry*. Paper presented at the 69th IFLA General Conference and Council, Berlin, Germany, August 1-9, 2003. Retrieved May 31, 2006 from http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Bar-Ilan, J., & Peritz, B. C. (2004). Evolution, continuity, and disappearance of documents on a specific topic on the web: a longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology*, 55(11), 980-990.

Brichford, M., & Maher, W. (1995). Archival issues in network electronic publications. *Library Trends*, 43(4), 701-712.

Brygfjeld, S. A. (2002). Access to Web archives: the Nordic Web Archive Access Project. *Zeitschrift für Bibliothekswesen und Bibliographie*, 49, 227-231.

CCSDS 650.0-B-1. (2002). Reference model for an Open Archival Information System (OAIS). Retrieved May 31, 2006 from Consultative Committee on Space Data Systems Web site: http://public.ccsds.org/publications/archive/650x0b1.pdf

Charlesworth, A. (2003). *A study of legal issues related to the preservation of Internet resources in the UK, EU, USA and Australia*. Retrieved May 31, 2006 from Joint Information Systems Committee Web site: http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf

Crichlow, R., Davies, S., & Wimbush, N. (2004). Accessibility and accuracy of Web page references in 5 major medical journals. *JAMA: the Journal of the American Medical Association*, 292(22), 2723-2724.

Dale, R. L. (2005). Making certification real: developing methodology for evaluating repository trustworthiness. *RLG DigiNews*, 9(5). Retrieved May 31, 2006 from http://www.rlg.org/en/page.php?Page_ID=20793

Darlington, J. (2003). PRONOM - a practical online compendium of file formats. *RLG DigiNews*, 7(5). Retrieved May 31, 2006 from http://www.rlg.org/preserv/diginews/diginews7-5.html

Day, M. (2001). Metadata for digital preservation: a review of recent developments. In P. Constantopoulos & I. Sølvberg (Eds.), *Research and advanced technology for digital libraries, 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001* (pp. 161-172). Lecture Notes in Computer Science, 2163. Berlin: Springer.

Day, M. (2003). *Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust*. Retrieved May 31, 2006 from Joint Information Systems Committee Web site: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

Day, M. (2004). Preservation metadata. In G. E. Gorman & D. G. Dorner (Eds.), *Metadata applications and management* (pp. 253-273). International Yearbook of Library and Information Management, 2003-2004. London: Facet.

Day, M. (2005). Metadata. In S. Ross & M. Day (Eds.), *DCC Digital Curation Manual*. Retrieved May 31, 2006 from Digital Curation Centre Web site: http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). Going, going, gone: lost Internet references. *Science*, 302, 787-788.

Digital Preservation Testbed. (2003). *Emulation: context and current status*. Retrieved May 31, 2006 from Nationaal Archief Web site: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf

Feeny, M. (1999). *Digital culture: maximising the nation's investment*. London: National Preservation Office.

Fitch, K. (2003). *Web site archiving: an approach to recording every materially different response produced by a Website*. Paper presented at the 9th Australasian World Wide Web Conference, AusWeb03, Sanctuary Cove, Queensland, Australia, July 5-9, 2003. Retrieved May 31, 2006 from http://ausweb.scu.edu.au/aw03/papers/fitch/

Garrett, J., & Waters, D. (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information*. Washington, D.C.: Commission on Preservation and Access; Mountain View, Calif.: Research Libraries Group. Retrieved May 31, 2006 from http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf

Giaretta, D., Rankin, S., McIlwrath, B., Rusbridge, A., & Patel, M. (2005) Representation Information for interoperability now and with the future. In *Local to global data interoperability - challenges and technologies, IEEE Mass Storage Systems & Technology Committee, Sardinia, Italy, June 20-24, 2005* (pp. 42-46). Piscataway, N.J.: Institute of Electrical and Electronics Engineers.

Gomes, D., & Silva, M. J. (2005). Characterising a national community Web. *ACM Transactions on Internet Technology*, 5(3), 508-531.

Hakala, J. (2004). Archiving the Web: European experiences. *Program*, 38(3), 176-183.

Hedstrom, M. (1998). Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, 31(3), 189-202.

Hedstrom, M. (2002). The digital preservation research agenda. In *The state of digital preservation: an international perspective* (pp. 32-37). Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub107abst.html

Hester, E. J., Heilig, L. F., Drake, A. L., Johnson, K. R., Vu, C. T., Schilling, L. M., & Dellavalle, R. P. (2004). Internet citations in oncology journals: a vanishing resource? *Journal of the National Cancer Institute*, 96(12), 969-971.

Hey, T., & Trefethen, A. (2003). The data deluge: an e-science perspective. In F. Berman, G. Fox & A. J. G. Hey (Eds.), *Grid computing: making the global infrastructure a reality* (pp. 809-824). Chichester: Wiley.

Hoeven, J. R. van der, Diessen, R. J. van, & Meer, K. van der. (2005). Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects. *Journal of Information Science*, 31(3), 196-208

Hunter, J., & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries*, 6(2), 174-183.

ISO 14721:2003: Space data and information transfer systems -- Open archival information system -- Reference model. Geneva: International Organization for Standardization.

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, 9(2), 174. Retrieved May 31, 2006 from http://informationr.net/ir/9-2/paper174.html

Koerbin, P. (2005). *Report on the crawl and harvest of the whole Australian Web domain undertaken during June and July 2005*. Retrieved May 31, 2006 from National Library of Australia Web site: http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å., Kruger, A., & Giles, C. L. (2001). Persistence of Web references in scientific research. *Computer*, 34(2), 26-31.

Lee, K. -H., Slattery, O., Lu, R., Tang, X., & McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology*, 107, 93-106.

López Borrull, A., & Oppenheim, C. (2004). Legal aspects of the Web. *Annual Review of Information Science and Technology*, 38, 483-548.

Lorie, R. A. (2002). *The UVC: a method for preserving digital documents*. Amsterdam: IBM Netherlands. Retrieved May 31, 2006 from Koninklijke Bibliotheek Web site: http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

Ludäscher, B., Marciano, R., & Moore, R. (2001) Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *SIGMOD Record*, 30(3), 54-63.

Lyman, P. (2002). Archiving the World Wide Web. In *Building a national strategy for digital preservation* (pp. 38-51). Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub106abst.html

Lynch, C. (1996). Integrity issues in electronic publishing. In R. P. Peek & G. B. Newby (Eds.), *Scholarly publishing: the electronic frontier* (pp. 133-145). Cambridge, Mass.: MIT Press.

Lynch, C. (1999). Canonicalisation: a fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, 5(9). Retrieved May 31, 2006 from http://www.dlib.org/dlib/september99/09lynch.html

Mellor, P., Wheatley, P., & Sergeant, D. (2002). Migration on request: a practical technique for digital preservation. In M. Agosti & C. Thanos (Eds.), *Research and advanced technology for digital libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002* (pp. 516-526). Lecture Notes in Computer Science, 2458. Berlin: Springer.

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., & Gupta, A. (2000) Collection-based persistent digital archives - part 1. *D-Lib Magazine*, 6(3). Retrieved May 31, 2006 from http://www.dlib.org/dlib/march00/moore/03moore-pt1.html

OCLC/RLG Working Group on Preservation Metadata. (2002). *A metadata framework to support the preservation of digital objects*. Dublin, Ohio: OCLC Online Computer Library Center. Retrieved May 31, 2006 from http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

PREMIS Working Group. (2005). *Data dictionary for preservation metadata* Dublin, Ohio: OCLC Online Computer Library Center. Retrieved May 31, 2006 from http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Rauch, C., & Rauber, A. (2004). Preserving digital media: towards a preservation solution evaluation metric. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. A. Fox & E. -P. Lim (Eds.), *Digital libraries: international collaboration and cross-fertilization, 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004* (pp. 203-212). Lecture Notes in Computer Science, 3334. Berlin: Springer.

RLG/OCLC Working Group on Digital Archive Attributes. (2002). *Trusted digital repositories: attributes and responsibilities*. Mountain View, Calif: Research Libraries Group. Retrieved May 31, 2006 from http://www.rlg.org/legacy/longterm/repositories.pdf

RLG-NARA Task Force on Digital Repository Certification. (2005). *An audit checklist for the certification of trusted digital repositories: draft for public comment*. Mountain View, Calif.: RLG. Retrieved May 31, 2006 from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Ross, S., & Gow, A. (1999). *Digital archaeology: rescuing neglected and damaged data resources*. London: South Bank University, Library Information Technology Centre.

Ross, S., & Hedstrom, M. (2005). Preservation research and sustainable digital libraries. *International Journal on Digital Libraries*, 5(4), 317-324.

Ross, S., & McHugh, A. (2005). Audit and certification of digital repositories: creating a mandate for the Digital Curation Centre (DCC). *RLG DigiNews*, 9(5). Retrieved May 31, 2006 from http://www.rlg.org/en/page.php?Page_ID=20793

Rothenberg, J. (1999). *Avoiding technological quicksand: finding a viable technical foundation for digital preservation*. Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub77.html

Rothenberg, J. (2000). *An experiment in using emulation to preserve digital publications*. Den Haag: Koninklijke Bibliotheek. Retrieved May 31, 2006 from http://nedlib.kb.nl/results/emulationpreservationreport.pdf

Sellitto, C. (2005). The impact of impermanent Web-located citations: a study of 123 scholarly conference publications. *Journal of the American Society of Information Science and Technology*, 56(7), 695-703.

Shepard, T. (1998). Universal Preservation Format (UPF): conceptual framework. *RLG DigiNews*, 2(6). Retrieved May 31, 2006 from http://www.rlg.org/preserv/diginews/diginews2-6.html

Smith, A. (2003). *New-model scholarship: how will it survive?* Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub114abst.html

Spinellis, D. (2003). The decay and failure of Web references. *Communications of the ACM*, 46(1), 71-77.

Strogatz, S. (2004). *Sync: the emerging science of spontaneous order*. London: Penguin.

Szalay, A., & Gray, J. (2006). Science in an exponential world. *Nature*, 440, 413-414.

Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. In *The state of digital preservation: an international perspective* (pp. 4-31). Washington, D.C.: Council on Library and Information Resources. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub107abst.html

Tibbo, H. R. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers*, 57, 1-67.

Van Bogart, J. W. C. (1995). *Magnetic tape storage and handling: a guide for libraries and archives*. Washington, D.C.: Commission on Preservation and Access; St. Paul, Minn.: National Media Laboratory. Retrieved May 31, 2006 from http://www.clir.org/pubs/abstract/pub54.html

Verdegem, R., & Slats, J. (2004). Practical experiences of the Dutch digital preservation test-bed. *VINE: the Journal of Information and Knowledge Management Systems*, 34(2), 56-65.

Waugh, A. (2006). The design of the VERS encapsulated object experience with an archival information package. *International Journal on Digital Libraries*, 6(2), 184-191.

Waugh, A., Wilkinson, R., Hills, B., & Dell'oro, J. (2000). Preserving digital information forever. In *ACM 2000 Digital Libraries, 5th ACM Conference on Digital Libraries, San Antonio, Texas, USA, June 2-7, 2000* (pp. 175-184). New York: Association for Computing Machinery.

Wren, J.D. (2004). 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, 20(5), 668-672.