Digital Preservation by Design.

A Chapter of the
Handbook of Research on Global Information Technology,
Mahesh S. Raisinghani (ed.)

By

Stacy Kowalczyk

Abstract

Current knowledge is produced, disseminated and stored in digital format. This data will not be preserved by benign neglect; digital information will be preserved only through active management. This chapter will provide a theoretical foundation for digital preservation, an overview of current best practice for digital preservation, and a research agenda as well as a proscriptive framework by which to design digital preservation into a system.

Introduction

An ever-increasing percentage of data, information, and expertise is being documented and distributed in an exclusively electronic medium. Our cultural history is now digital. Not only are family histories now being created with digital cameras, almost all current events are captured digitally. Newspapers, magazines, and books are produced digitally. Virtually all published photographs are created digitally. 35mm film is nearly extinct. All broadcast media is digital. Radio is recorded digitally – audiotape is nearly extinct. Television is produced and delivered digitally. Movies will be distributed digitally within the next few years. Not just cultural history, but predictions are that scholarly output will be exclusively digital by 2010.

Businesses, as producers of knowledge, create many documents that contribute both to our cultural history and to the scientific record. Contracts, white papers, marketing and promotional materials, audio and video documentaries of important occasions and people in the company are important not only to the company but to the world. All types of research from pharmaceutical and medical to electronics and computer science to new power technologies are business activities – and all digital. From buildings to clothing, almost all design in digital. All commerce is digital. *Current knowledge is produced, disseminated and stored in digital format.*

What is the problem? Isn't information in digital form more secure than paper? It is not a mater of degree. It is a matter of understanding the differences between a fixed media and digital. While paper can be put in a box on a shelf, digital objects are increasingly expensive to store because the media needs to be replaced on a three to five year schedule. Unlike paper which will remain useable if put in a relatively safe place, digital objects require constant and perpetual maintenance. Digital objects depend on and are bound to a technical environment and infrastructure. As the environment changes, so might the objects. And as the digital objects become more complex, the problems just increase. *Data will not be preserved by benign neglect. Digital materials will only be preserved by active management.*

Because so much information exists only in digital format, concern about our ability to keep this information available and usable for the future is increasing. In August, 2006, a prime example of a digital disaster made headlines. Tapes with very high quality pictures of the first moon walk in 1969 have been reported as missing. To add insult to injury, the only remaining tape drive that can read these tapes is scheduled to be retired within the year. The missing tapes have a much higher quality picture than what was seen on, and recorded for, television (Macey, 2006). The loss is incalculable.

Businesses, governmental agencies, libraries, archives, and museums all need solutions to these problems. Over the past 10 years, digital preservation has emerged as an important area of research in both computer science and information science, at a national and international scale. Much of the research in digital preservation is coming from digital library programs more than from the traditional academic research community. Digital preservation is defined as “the managed activities necessary: 1) For the long-term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology” (RLG & OCLC, 2002).

This chapter will provide a theoretical foundation for digital preservation, an overview of current best practice for digital preservation, and a research agenda.

Digital Preservation Strategies

In a research area that is only 10 years old, it seems almost silly to talk of the "early days"; but in the early days, the focus of the conversation and theorizing revolved around preservation strategies. Three strategies emerged: technology preservation, technology emulation and data migration. These are best defined though example. We will use the trivial example of a CD-ROM application that provides a virtual tour of the Vatican.

The technology preservation strategy proposes to save the technology platform for an application to preserve not only the data, but the "look and feel" as well. In our example, we would need to save the CD-ROM, a computer with a CD-ROM reader with the appropriate read speed as well as the correct operating system for the CD-ROM application software. In order to take the virtual tour, one would need to use that specific computer. Known as the museum style approach (UKOLN, 2006), one can see the attraction to this model – it seems simple to just keep the hardware functioning. However, its simplicity is also its downfall. Depending on the number of applications to be kept functioning and the number of different platforms, the complexity will escalate until the cost becomes prohibitive. But the most significant failure point is that eventually, due to the increasing age, the hardware and storage media will fail irrevocably. Preserving the hardware as a scientific artifact for the future is a worthy goal. A number of museums are collecting computers. The Science Museum and the Computer Conservation Society in the UK are primary examples (UKOLN, 2006). But as a preservation strategy, because of its obvious weaknesses, it was never considered to be a serious option.

Like technology preservation, technology emulation is a strategy that has as its goal the preservation of the complete functionality of the original system. This strategy is based on the premise that system behaviors can be defined independently from its implementation. With the description of behavior, an engine could be developed that would re-create those behaviors. Alternatively, the source could be captured and saved and used as input into an emulation engine. As computers become faster and cheaper, the processing costs to deliver the information through an emulator would be negligible (Rothenberg, 1999a; Rothenberg, 1999b; UKOLN, 2006). In our example, the CD-ROM would be copied in its entirety onto an emulation server. The server would need to know which operating system to emulate and how to invoke the software as well as the wide variety of peripheral hardware interfaces. The server would also need to know what the underlying formats of the data, images, text, and video in order to render them to a client application. While the technology preservation strategy was quietly dismissed as a viable option early in the discussion, emulation caused a firestorm of controversy (Bearman, 1999; Fleischhauer, 2003). Most critics cited the overwhelmingly daunting task of maintaining the requisite knowledge of a wide variety of hardware technologies, operating systems, database, data formats, and other operating environments, as well as the functionality of each of the applications to be preserved. What was not widely discussed at the time, but which as become more obvious over time, is the confusion of presenting the data to consumers of the information preserved. How will future generations deal with command-level 1960's mainframe applications or 1980's game interfaces?

The data migration strategy proposes to move data into new formats as the old formats and/or systems become obsolete (Waters & Garret, 1996). The goal of this strategy is to preserve the data – the actual information – with as much of its functionality as possible. In our example, the data would be extracted from the CD-ROM. The text would be formatted into a useable format (today that would be an XML-based text file) with links (today that would be a URL-based link) to the images (today would be TIFF as a master file with a JPG deliverable file). A new system would need to be created that would be able to render a variety of XML-based text files. People using this new application would see the same data, but in a different environment with a different set of actions. Perhaps some specific functionality of the old system – perhaps an animated travel companion – would be lost. The emulation proponents decry the loss of functionality that the proponents dismiss as trivial. What is universally considered to be a drawback to the migration strategy is the loss of data as a result of the migration itself. Moving from technical format to another often results in loss. If the image files were in the now obsolete PhotoCD format, migrating them to TIFF would result in irrevocable data loss. Even with these limitations, data migration is the predominate strategy to date.

OAIS Reference Model

The Open Archival Information Systems (OAIS) Reference Model has become a foundation for discussing digital preservation. Like the OSI reference model, it is not a systems design but a set of high-level requirements built as a conceptual framework. It lays out what should be done in a digital preservation archive. It does not provide implementation instructions. OAIS uses the concept of an Information Package, a transaction and/or data store that accompanies a digital object. An information package is a set of metadata that accompanies a digital object. OAIS uses the traditional meaning of metadata – data about (or describing) data. When data is brought into the system, the package is a SIP, a submission information package. It is modified as required to be come an AIP, an archival information package. When the object is distributed for use, the information package is transformed again into the DIP, the dissemination information package (CCSDS, 2002). The OAIS was developed by the Consultative Committee for Space Data Systems. The OAIS conceptual model has made two major contributions to the discussion of digital preservation. The first is that concept that preservation is a planned process that needs to be designed into a digital library from inception. The second is a common language for discussion digital preservation from submission to dissemination, from ingest to administration.

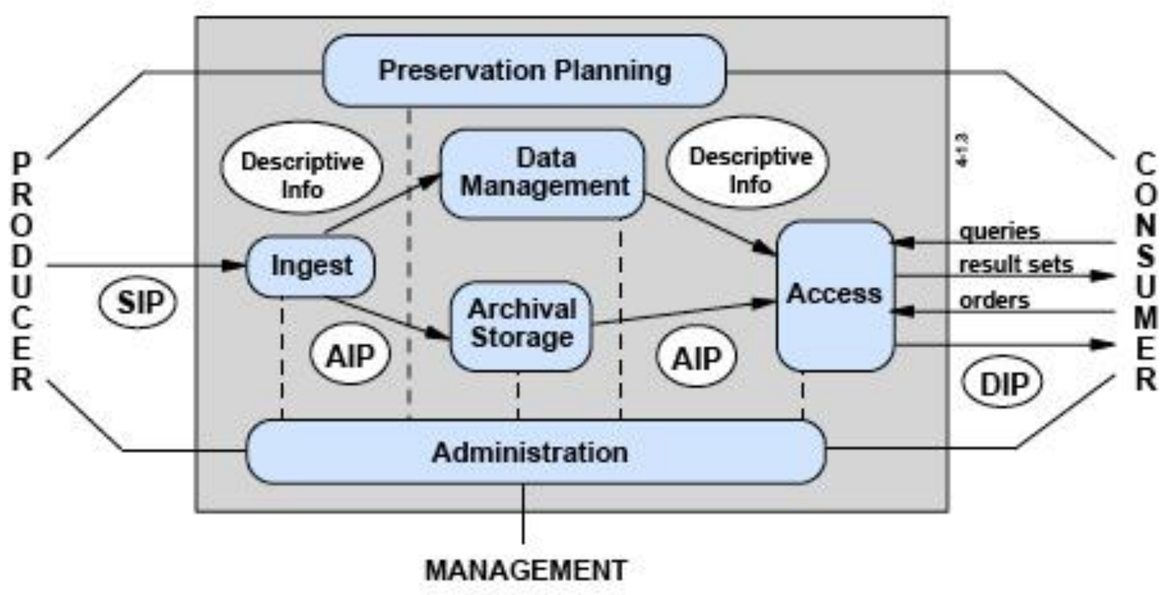


Figure 4-1: OAIS Functional Entities

OAIS Reference Model Overview (CCSDS, 2002)

Used with permission

In the seminal paper on digital preservation, Waters and Garrett state:

> For digital object, no less than for object of other kinds, knowing *how* operationally to preserve them depends, at least in part, on being able to discriminate the essential features of *what* needs to be preserved…Whatever preservation method is applied, however, the central goal must be to preserve information integrity: that is, to define and preserve those features of an information object that distinguish it as a whole and singular work… including: content, fixity, reference, provenance and context. (1996).

The 4 Goals of Preservation

Is this not a solved problem? Many think that if multiple copies exist in multiple locations, if the storage media is monitored for data degradation or technological obsolescence, if data center best practice that has been developed over the past 40 years is followed, then the data is "preserved."

Unfortunately, "keeping the bits safe" is only the first, and the easiest, step in digital preservation. The longevity of digital information depends on more than good backups. Waters and Garrett outlined four major goals of digital preservation – keeping the bits safe, keeping the files useable, keeping the integrity of the object, keeping the context of the object (1996). Fulfilling these goals will produce a trusted digital object which can prove the authenticity of its underlying bit-stream (Gladney, 2004). If we can design systems with preservation in mind, we can build trustworthy repositories for trusted digital objects.

Keeping the Bits Safe

Insuring the integrity of the bit-stream is the most basic function of digital preservation. Over the past 40 years, a body of best practice has been developed to preserve data at the bit-stream level. Fortunately, most businesses have implemented data center best practices. Best practice dictates a minimum of three copies of any file or record; one on active media; one on a near-line backup media and one on a remote backup media. Unfortunately, creating the copies is not sufficient. One must manage these files. The media needs to be rotated and refreshed based on manufactures most conservative lifetime estimates. Media reliability degrades with the number of writes. Backup tapes or other removable media need to be rotated off the backup schedule after a certain number of uses. Offsite media need to be refreshed on a regular schedule. After some number of years (again, based on the manufacture's recommendation), the offsite media need to be rewritten. But this schedule also needs to be managed. Will the organization have the necessary technology to rewrite the tapes? Thus, technology needs to be monitored to reduce the risk of obsolesce. When storage media is replaced, any removable media that uses that infrastructure must be replaced. This often greatly increases the cost and the complexity of a technology replacement project.

Storage

Storage – specifically choosing the technology – is always a concern as organizations begin to design a digital repository. As the Assistant Director for Software Development in the Harvard University Library Digital Initiative, the first question that I was asked when I spoke about digital repositories was "What storage hardware are you using?" As I did then, I would like to discuss storage in the abstract – what issues need to be analyzed during the hardware selection process. The National Archives of England, Wales and the United Kingdom has developed an excellent set of selection criteria and a media scorecard in their second Digital Preservation Guidance Note. The six criteria are as follows:

- *Longevity* – is not as important as one might expect since most storage media becomes obsolete before it degrades. The National Archives of England, Wales and the United Kingdom recommends a "proven life-span" of 10 years (Brown, A., 2003b p. 5).
- *Capacity* – is actually more important than longevity. Limiting the number of devices to manage is the key. Estimating the amount of storage required is the first and most difficult step.
- *Viability* – describes the hardware level data protection features of the storage media. Can the media be write protected? What types of read/write error detection is available? Is the media able to self-recover at failure?
- *Obsolescence* – is, of course, a major concern. How does one decide between leading-edge and mature? We want to invest our limited resources in a technology that is both stable and has a high "cool" factor. Choosing technologies built on open standards with a high level of interoperability has a significantly higher probability of being usable longer.
- *Cost* – has two major considerations – initial cost and total cost of ownership. The cost per gigabyte, the Mean Time Before Failure, as well as the personnel cost of administration should be factored into the total cost equation.

- *Susceptibility* – is the term that describes the media's ability to withstand physical damage. Any media used should "be tolerant of a wide range of environmental conditions without data loss" (Brown, A., 2003b p. 5).

Creating an expandable and extendable storage architecture based on new but proven technologies, while seemingly more expensive, is more efficient to manage and highly likely to have a lower total cost of ownership.

The OAIS Archival Storage Model visually demonstrates that developing an architecture for preservation requires more than hardware. It requires process and management.

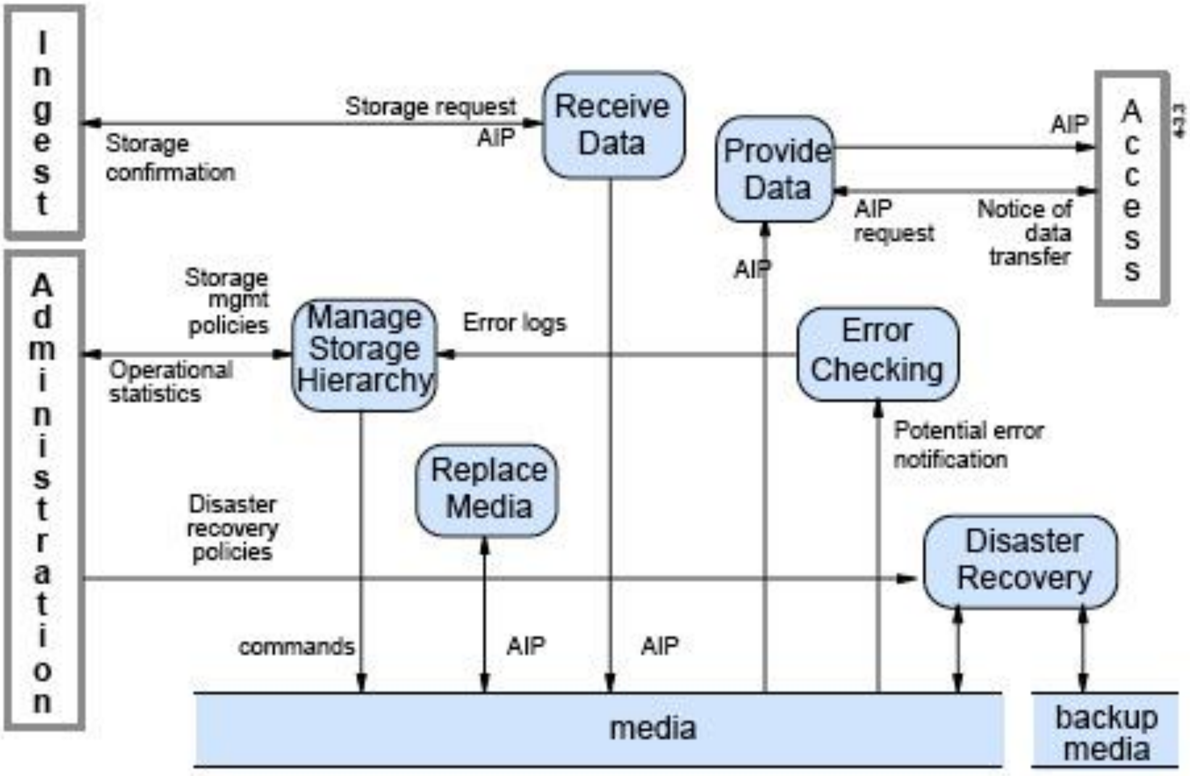


Figure 4-3: Functions of Archival Storage

OAIS Archival Storage Model (CCSDS, 2002)

Used with permission

Archiving

Besides the body of best practice developed for data centers, more must be done to create the Trusted Digital Object that Gladney proposed (2004). As work becomes more collaborative and more distributed, organizations find it increasingly difficult to manage their complex data. As a business develops new products, consultants in the US develop business plans on a Microsoft platform; employees in Europe create the designs using CAD software; and the manufacturing plant in China uses proprietary systems to build the prototypes. While people are actively working on a project, information is emailed and documents are stored on shared file system. When this company wants to harvest the knowledge and create a digital archive of the work products of the development, rather than a working dataset or file system, an archival digital object needs to be created. It then becomes the responsibility of the application to insure the integrity of the object both its fixity and its validity at inception and over time.

Fixity is the term used to convey the notion that digital data is not immutable, that it can be changed easily, either maliciously or inadvertently. Regardless of intention, data cannot be considered to be preserved as a Trusted Digital Object if the repository cannot guarantee fixity – that the data has not been changed since it was archived. Technically, this is an insignificant task, but it does require a process to enforce. Each archived object needs to have a checksum or a digital signature. Because it is both simple and effective, many digital libraries use a simple MD5 checksum rather than the more cumbersome digital certificate. This checksum is then stored in a database with administrative data about the object. But a checksum is only useful if it is validated by a process that calculates the checksum for each digital file and compares it to the saved checksum. If they do not match, an error needs to be sent to the appropriate staff for immediate action. Errors can be attributed to a number of factors – failing disk, ingest errors, file transfer errors, as well as malicious mischief.

To insure fixity, a digital repository should implement a checksum or digital signature on archived files and validate them on a regular schedule.

IT Contingency Planning

Disaster recovery planning, business resumption planning and IT contingency planning are considered to be synonymous terms. But regardless of the term, it is a process, which allows the administrators of a system to know the tolerance for down time and provide a guide for restoring service under a wide variety of circumstance.

The National Institute of Standards and Technology (NIST) has developed a standardized model for developing a contingency plan:

1. Develop the contingency planning policy statement
2. Conduct the business impact analysis (BIA)
3. Identify preventive controls
4. Develop recovery strategies
5. Develop an IT contingency plan
6. Plan testing, training, and exercises
7. Plan maintenance. (NIST, 2002)

As is obvious from the NIST process described above, contingency planning is complicated. While the discussion could include hardware and network redundancy, automatic failover and site mirroring, for the purposes of this chapter, disaster recovery planning will be limited to the issues of data preservation only.

To insure that a digital repository can "keep the bits safe", the data must not only be safe on the primary disk and backup media. The administrators of the system must be able to find the correct version, have access to the restore software, and have experience of restoring data. This requires an annual disaster recovery drill, which involves actually finding and retrieving data file, restoring the file and testing to insure data integrity. This should be incorporated into a larger disaster recovery and business resumption annual drill. ***A digital repository should institute an annual contingency plan drill within 6 months of its initial production date***.

Keeping the Files Usable

Of the four digital preservation goals established in 1996 (Waters & Garret), keeping the files usable has turned out to be the most challenging because it depends on the complexity and transparency of file properties. Not only do digital objects depend on technology, they are bound to an environment and infrastructure. Changes to the environment or infrastructure might the objects. Abrams contends that "the concept of representation format permeates all technical aspects of digital repository architecture and is, therefore, the foundation of many, if not all, digital preservation activities" (2004). In other words, the technical format of a file determines its probability of being preserved. A format is defined as "the internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in a human-accessible form" (Brown, A., 2006 p. 3).

Formats need to be discussed within the context of the "levels of use." In digital preservation, we expect digital objects to be stored in different formats for different uses. In order to keep a file useable for the longest period of time, a digital object should be created in the format with the most information, in the most open format with the least risk of failure. The highest quality object with the highest level of fidelity is called an archival master file. Derivative files are made from the master file as necessary for more efficient delivery to applications. For images, current best practice would dictate that the master file be an uncompressed TIFF file and that images to be sent to a web browser be a JPEG file. For audio file, current best practice would dictate that the master file be an uncompressed AIFF or Broadcast WAV and that the files to be sent to a web browser be an MP3 or a RealAudio file (Library of Congress, 2006b; TASI, 2006; National Library of Australia, n.d.b).

Format Risk Assessment

How do organizations analyze the risk for an archival file format? Early projects, not surprisingly, discovered that the more open the format, the easier the process. But, they also discovered that even open formats can have a proprietary and thus, secret, set of tags that may not be supported in future version or have any documentation for migration or emulation work (Lawrence, Kehoe, Rieger, Walters & Kenney, 2000). Since the earliest projects, a number of organizations have developed different types of risk analysis.

The Library of Congress has developed a theoretical framework for assessing formats using seven sustainability factors.

1. ***Disclosure***. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.
2. ***Adoption***. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.
3. ***Transparency***. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.
4. ***Self-documentation.*** Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.
5. ***External Dependencies***. Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.

6. ***Impact of Patents***. Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.

7. ***Technical Protection Mechanisms.*** Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository (Arms & Fleischhauer, 2005).

The National Archives of England, Wales and the United Kingdom has an alternative list of seven criteria by which formats could be judged: open standards, ubiquity, stability, metadata support, feature set, interoperability and viability (Brown, A., 2003a). But both sets of criteria are aiming at a similar result – helping managers determine how to store their data for the long term. Archival formats should be:

- well documented, well understood and not wholly owned by a single commercial entity
- widely adopted to increase the probability of commercial tools for migration and to ensure a long usage cycle to avoid repeated, short term migrations
- be self-contained and not rely on a specific technical environment

But beyond just format analysis, organizations need to be aware of other risks in the preservation process. The National Archives of Australia has developed a process called DIRKS –Designing and Implementing Recordkeeping System. This is an eight-step process defining the best practice standards and guidelines published by the National Archives as part of its efforts to preserve electronic information (they use the term “e-permanence”). The DIRKS methodology is a basic waterfall requirements process that is designed to “help each organisation [sic] to determine such requirements and put in place procedures to reassess these needs over time” (National Archives of Australia 2003b). The goal of the DIRKS methodology is a set of records that are authentic, reliable, complete and unaltered, useable, with integrity (National Archives of Australia 2003b).

The multi-national library utility, OCLC, has developed the INFORM methodology for assessing the durability of formats for digital preservation which is based on 6 classes of risk that include both technology and organizational process:

1. ***Digital object format*** - risks introduced by the specification itself, but also including compression algorithms, proprietary (closed) vs. open formats, DRM (copy protection), encryption, digital signatures.
2. ***Software*** - risks introduced by necessary software components such as operating systems, applications, library dependencies, archive implementations, migration programs, implementations of compression algorithms, encryption and digital signatures.
3. ***Hardware*** - risks introduced by necessary hardware components including type of media (CD, DVD, magnetic disk, tape, WORM), CPU, I/O cards, peripherals.
4. ***Associated organizations*** - risks related to the organizations supporting in some fashion the classes identified above, including the archive, beneficiary community, content owners, vendors, open source community.
5. ***Digital archive* -** risks introduced by the digital archive itself (i.e., architecture, processes, organizational structures).
6. ***Migration and derivative-based preservation plans*** - risks introduced by the migration process itself, not covered in any other category. (Stanescu, 2004 p. 3)

Format Registries

In one of the early investigations into file format risk analysis, the authors indicate that biggest challenge in their research was compiling complete data about each of the formats under consideration (Lawrence et al., 2000). As other preservation-by-migration projects began, the problem of adequate format and persistent information became more obvious. In response, several institutions have initiated projects to create repositories of format information. Over the past several years, these have become known as "format registries." One of the major issues of format registries is the representation of information for format data, compression methods, character encoding schemes, and operating systems. The need for format registries is obvious when one looks at the general categories of formats that the National Library of Australia has outlined on their very useful online resource, Preserving Access to Digital Information: audio and audiovisual material, computer aided design (CAD), databases, digital television, digital games, email, geographic information system (GIS), networked digital material, physical format digital material, spreadsheets, virtual reality, word processing documents (National Library of Australia, n.d.b).

The National Archives of England, Wales and the United Kingdom has developed a format registry called PRONOM which provides information about data file formats and their supporting software products. Initially released in 2002, the system was developed for use by the National Archives but has recently been made available via the web for public consumption. PRONOM holds information about software products as well as the file formats which each product can read and write. Currently, PRONOM only represents format information. The most interesting feature of the PRONOM system is the PRONOM Persistent Unique Identifier (PUID). The PUID is a specific instantiation of the more generalized construct, the Persistent Identifier (PID) that will be discussed later in this chapter. In the specific case of the PRONOM system, the PUID is a persistent, unique and unambiguous identifier for formats within the PRONOM registry. "Such identifiers are fundamental to the exchange and management of digital objects, by allowing human or automated user agents to unambiguously identify, and share that identification of, the representation information required to support access to an object.

This is a virtue both of the inherent uniqueness of the identifier, and of its binding to a definitive description of the representation information in a registry such as PRONOM" (National Archives of England, Wales and the United Kingdom, 2006b). The PUID is being used in the latest version of the e-Government Metadata Standard – providing a consistent method for describing file formats used throughout the UK government. While PIUDs can be expressed as Uniform Resource Identifiers (URIs) using the "info:pronom/" namespace, they currently can not be resolved to a URL, although the National Archives plans to implement this in the future (National Archives of England, Wales and the United Kingdom, 2006b).

The Digital Library Foundation (DLF) has sponsored an initial investigation into the creation of a global digital format registry (GDFR) to maintain format representation information. One of the most important contributions of this project is its international participation which includes the Bibliothque Nationale de France, Harvard University, the Joint Information Systems Committee of the Higher and Further Education Councils in the United Kingdom (JISC), JSTOR, the Library of Congress, Massachusetts Institute of Technology, the National Archives and Records Administration, the National Archives of Canada, the National Institute of Standards and Technology, New York University, the Online Computer Library Center, the University of Pennsylvania, Stanford University, the British Library, the California Digital Library, the Internet Architecture Board, the Internet Engineering Task Force, the Research Libraries Group, and the Public Records Office in the United Kingdom (Digital Library Federation, n.d.). The project is working on a long-term business model for sustaining the operation of the registry (Abrams, 2004). A prototype system is being developed by the University of Pennsylvania called FRED – Format REgistry Demonstration (Ockerbloom, 2004). This prototype is built on TOM – the Typed Object Model. TOM is both an abstracted data model that describes the behaviors and representations of information sources such as file formats and information retrieval services and a set of software services that use the model (Ockerbloom, 2005).

The Digital Curation Centre (DCC) is an organization that is developing another repository. DCC is supported by the Joint Information Systems Committee (JISC) of the United Kingdom. Using the OAIS conceptual model that was described in the opening section of the chapter and will be further explicated in the following section, the DCC is implementing what OAIS describes as a representation information registry/repository for digital data (Digital Curation Centre, 2006). The DCC has architected the system for general applications of digital curation – format obsolescence – but is primary interested in experimental scientific data which has significantly different properties than regular file formats and thus has more complex functional requirements than the format registries described above. As a minimum, the registry/repository needs format specifications, the details of the bit structure. But it also needs rendering and processing software source code as well as binary executables along with an extensive set of metadata that is often considered data provenance, which will be discussed later in the chapter.

Redaction

There are many circumstances in which an organization will want to store a complete archival digital object that contains sensitive data while needing to make the nonsensitive data available to their employees or the general public. The process of removing content is common occurrence in records management and is referred to as redaction. Redaction is “the separation of disclosable form non-disclosable information by blocking out individual works, sentences or paragraphs or the removal of whole pages or section prior to the release of the document” (National Archives of England, Wales and the United Kingdom, 2006a p. 4). The National Archives of England, Wales and the United Kingdom has developed a set of five principles for redacting electronic records.

- The original file must not be altered. Redaction should be performed on a copy of the record.
- Redaction should "irreversibly remove the required information for the redacted copy of the record" (p 13) not merely obstruct display.
- Redaction techniques should be fully tested for security
- Redaction should be done in a controlled and secure environment
- All intermediary states of the process should be deleted. (2006a)

While redaction may be necessary, it is not recommended as a general practice on archived digital objects.

Data Ingest

Data ingest, as the name implies, is the process of getting data into a system. Best practice for all information systems is to get the data "right" right from the beginning. Fixing errors in data is always expensive. But for digital repositories, it is absolutely essential to get the correct digital file, intellectual metadata, and technical metadata at time of ingestion (National Archives of Australia 2003a). According to the National Archives of England, Wales and the United Kingdom, the cost of creating data for sustainability should be the goal because "attempts to bring electronic records into a managed and sustainable regime after the fact tend to be expensive, complex and, generally, less successful" (Brown, A., 2003a p 4).

In the OAIS model, ingest is the initial stage of preservation. When data is deposited in the repository, it must be accompanied by a Submission Information Package (SIP). The SIP must have sufficient information about the object to be processed. The repository must confirm the quality of the data – both from the SIP and from the digital object itself. Upon successful ingest, the system must create an Archival Information Package (AIP) to insure that sufficient data exists within the system to preserve the object. All data is stored and is actively managed over time.

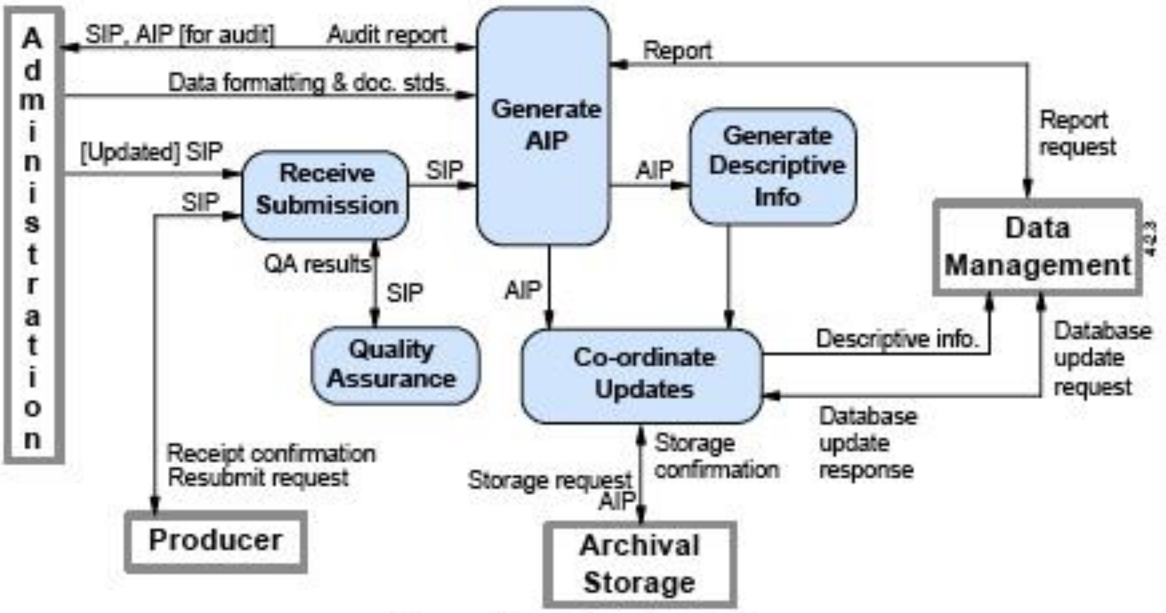


Figure 4-2: Functions of Ingest

OAIS Ingest Model (CCSDS, 2002)

Used with permission

National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress is a 10 year, multi-million dollar project to develop a distributed network of both public and private preservation partners. Ingest is a vital component of this initiative; the Archive Ingest and Handling Test (AIHT) is an NDIIPP project to text the ability of 8 partner institutions to test the process of ingesting a common set of digital objects and later export and exchange with another partner. The major issues were not technical but developing a shared language and set of values (Smith, 2006).

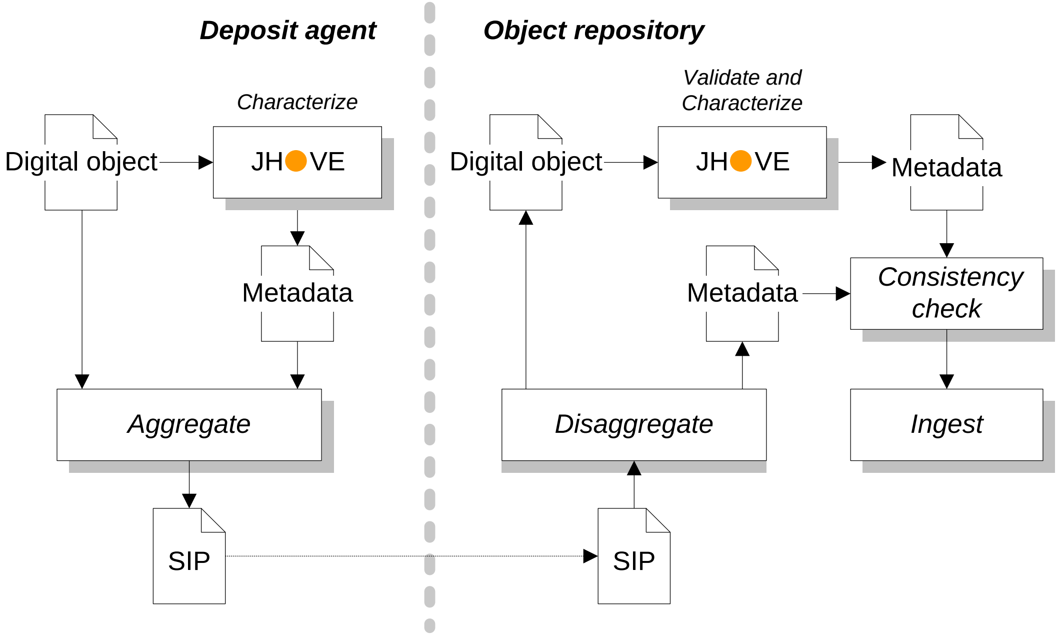
Identification and Validation

We have already established the primacy of technical format in digital preservation. Ensuring that the digital repository knows with certainty that the file is what it purports to be and that the file conforms to all of the specifications of the format is vital. With the volume of objects that organizations need to process at ingest, an automated process to both identity and validate the format is required.

Identification is the first step. Most files self-identity via the file extension – .txt indicates a text file; .doc indicates a Microsoft Word file, .tif indicates a TIFF file; .jpg indicates a JPEG file; the list could continue for pages. But self-identification is unreliable and insufficient for a digital repository. Files are not always in the format that they claim to be. "Invalid objects can arise through the use of poor-quality software tools or as the result of accidental or deliberate corruption" (Brown, A., 2006 p 4). Not only does the repository needs to know with certainty that the file is a TIFF file, it also needs to know that the file was created using the TIFF 6.1 specification.

Validation is the processes of determining the level of conformance to the encoding specification of a specific format. This process has two steps – the first step is to determine if the object is well-formed. Well-formed objects conform to the syntactic requirements for its format. In other words, it follows the grammar of the format. The second step is to determine its validity. Valid objects must conform to the semantic requirements of the format – does the object contain the meaningful content that is required. If a TIFF file has an 8-byte header followed by a series of Image File Directories (IFDs) made up of a two byte entry count and a set of 8 byte tagged entries, it can be considered to be well-formed. But to be considered valid, it must conform to additional rules that enforce more complex semantic rules – an RGB TIFF must have at least three sample vales per pixel (Harvard University Library, 2006).

While it might be possible to identify many file types via the UNIX file command using the "magic number" (Brown, G. 2006), many feel that this is insufficient (Abrams & Seaman, 2003). Harvard University Library and JSTOR, a not-for-profit organization that creates and maintains a trusted archive of important scholarly journals, have collaborated through a Mellon funded project to create a tool set to automate format-specific validation of digital objects. The software is named JHOVE – the JSTOR/Harvard Object Validation Environment (Harvard University Library, 2006). In addition to identification and validation, JHOVE also allows for characterization – the processes of determining the format-specific significant properties of an object. These actions are performed by modules which plug into a layered architecture that can be configured at the time of its invocation to include specific format modules and output handlers. JHOVE includes modules for arbitrary byte streams, ASCII and UTF-8 encoded text, GIF, JPEG2000, and JPEG, and TIFF images, AIFF and WAVE audio, PDF, HTML, and XML; and text and XML output handlers. JHOVE has been used by a number of digital libraries and archives. As of the writing of the chapter, JSTOR and Harvard were working the Library of Congress to enhance the functionality and to extend the architecture to improve the performance and the interoperability of the system (Harvard University Library, 2006).



JHOVE Schematic (Harvard University Library, 2006)

Used with the permission of the President and Fellows of Harvard College

The National Archives of England, Wales and the United Kingdom has developed an alternative method for identifying formats. The approach taken by The National Archives is to develop a format "signature." Using information about each specific format stored in the PRONOM format registry described above, the Digital Record Object Identification system analysis the binary structure of a digital object and compares it with this predefined signature.

The National Archives expects to extend this system to include format validation (Brown, A., 2006).

A digital repository should optimize the probability of preservation by limiting the number formats accepted as archival quality or by reformatting the data upon ingest. A digital repository should validate digital objects when submitted for ingest. The validation should include format identification, validation and characterization.

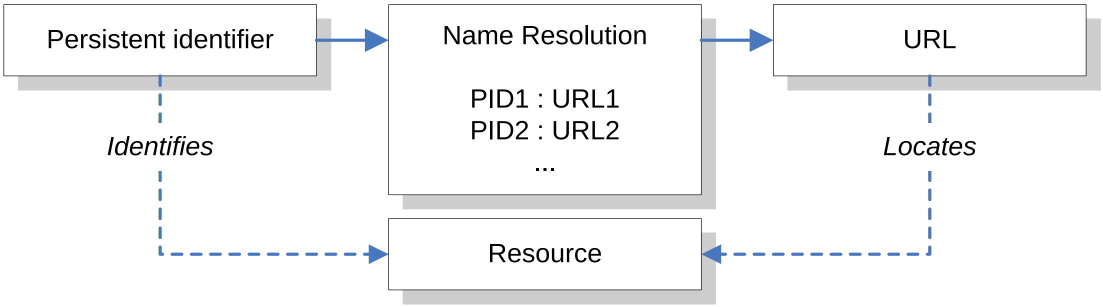
Data Provenance

Data provenance is a term that "broadly refer[s] to a description of the origins of a piece of data the process by which it arrived in a database" (Buneman, Khanna & Tan, 2000, p 2). Also know as lineage, provenance is a "special form of audit trail that traces each step in sourcing, moving and processing data" (Pearson, n.d.). In a repository, whether a generalized digital repository or a domain-specific repository, data provenance is difficult. Defining the data required to prove provenance is specific to each data type and/or domain. Developing these ontologies is a "crucial and thankless" task (Saltz, n.d.). Domain specific provenance schemas are being developed – SNOMED and LOINC for medical research (Saltz, n.d.) and Karma for atmospheric research (Plale, Ramachandran & Tanner, 2006; Simmhan, Plale & Gannon, 2006) are but a few of the research efforts. But what is needed is a more generalized solution. Providing self-describing data and the software to process that data is not yet a reality. Chimera, a prototype generalized data provenance system, has attempted to abstract the data representation by using types, descriptors and transforms. Types provide an abstraction layer for semantic content, the format of the physical representation and the format's encoding; descriptors define an interpretable schema that defines the data – number of files, types of files, etc; and transforms are a generic, typed abstraction for computational procedures (Foster, 2002).

While much of the research on data provenance is in the scientific community, the issues surrounding data provenance permeates digital preservation and digital repositories. Unless repositories can fully describe how files have been transformed, either for research or for preservation, the objects will not be trustworthy. ***A digital repository should provide provenance information for all of its digital objects.***

Persistent Identifiers

A persistent identifier (PID) is a unique, permanent, location-independent identifier for a network-accessible resource. A URL cannot be a persistent identifier because it conflates two important but separate functions – item location and item identification. By separating location from identification, PIDs help avoid the problem of broken links that is often referred to in the digital library community as the “404 – not found problem”. Persistent identification is especially crucial for important resources that are referenced from scores of web pages or are cited scholarly works. Basically, PIDs are URLs that require a multi-phased resolution process. The first level resolution is to find the right resolution server. Once the resolution server is located, the second level resolution finds the resource. The second phase of the resolution is relatively simple. A database has both the persistent id and a location of the object usually in the form of a URL. Using the PID as a key, the resolution service reads the database and redirects the http transaction using the location URL. As the problem of persistent identification of digital resources has become more prominent, a number of competing PIDs have been developed.



Harvard University’s Name Resolution Service (Harvard University Library, 2003b)

Used with the permission of the President and Fellows of Harvard College

The Harvard University Library developed an early implementation of a persistent identifier system called the Name Resolution Service (NRS). This service is based on the Universal Resource Name (URN) syntax (WC3, 2001) with the expectation that web browsers would ultimately support URNs as specified. But until that native browser support, the URN system needed to be delivered in URL syntax (Harvard University Library, 2003b). The Online Computer Library Center (OCLC) uses a PURL – persistent URL – a technology similar to the Harvard method but with a different syntax (Weibel, Jul, & Shafer, 1995). The Corporation for National Research Initiatives (CNRI) created another resolution system for its unique persistent identifier syntax named Handles. The Handle system is an open source, downloadable system with a set of open protocols and a namespace (Corporation for National Research Initiatives, 2006). The International DOI Foundation developed the Digital Object Identifier (DOI) naming semantics that has been implemented using the Handle syntax and resolution system (International DOI Foundation, 2006). DOIs are used predominantly by publishers of electronic information to provide persistent and controlled access to journal articles. The California Digital Library has developed the Archival Resource Key (ARK) which provides access to a digital information resource. The ARK is architected to deliver three parameterized services – the digital object metadata, digital object content files, and a commitment statement made by the owning organization concerning the digital object (Kunze, 2003).

A digital repository should implement a persistent identifier service to insure long-term unambiguous access to its digital objects.

Keeping the Integrity of the Object

More than fixity, the integrity of the object has referred to maintaining the intellectual wholeness of a digital object. Remember that a digital object is one or more files that constitute a logical entity. In a scientific application, a digital object could be several ASCII datasets with XML documentation and 10 visualizations stored as jpg images as well as an XML workflow process file. In a medical application, it could be a series of x-rays, a treatment plan, a set of diagnostics, and a patient visit chart. For a digital book, it is a series of page images, OCR text files, a word/page coordinate file for highlighting a search hit on a displayed image, a descriptive metadata record with author and title information. Examples of complex digital objects seem limitless.

Over the past years, digital libraries have grappled with the issue of managing these complex objects. Initially, file-naming conventions were used to associate all of the files. That proved to be unsatisfactory and insufficient. After testing several different schemes, libraries developed a Metadata Encoding and Transmission Standard (METS) XML schema for maintaining the complex relationships between different files. A METS document consists of seven major sections:

1. **METS Header** - The METS Header contains metadata describing the METS document itself, including such information as creator, editor, etc.
2. **Descriptive Metadata** – Access metadata can be stored internally or have an external link or both.
3. **Administrative Metadata** – Provenance and other administrative data
4. **File Section** - Lists all files containing content which comprise the object.
5. **Structural Map** – Provides a hierarchical structure for the digital object and links the elements of that structure to content files and metadata that pertain to each element.
6. **Structural Links** - Records the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map.
7. **Behavior** - Associates executable behaviors with content in the METS object. (Library of Congress, 2006a)

Some digital libraries use METS internally; others use an RDBMS internally but export METS. But METS, or any other metadata schema, is not sufficient for insuring the integrity of an object. Many of the systems that manage digital objects, to be covered in detail in the following section, have a similar architecture – the data resides on a file system, and administrative and intellectual metadata reside in a database. While early attempts were made to keep the digital files in a database, this proved to be impractical. The huge file sizes caused the database to expand so rapidly that the database could not be backed up in a regular batch window. Other solutions, either database or hardware storage architectures, proved very costly. Thus the two tiered architecture. How can a repository insure object integrity with data residing in different systems?

The Harvard University Library's Digital Repository System (DRS) has developed a set of procedures that insure the integrity of their digital objects. The DRS uses METS to keep intellectual control over the digital objects. When a complex object is ingested, the DRS creates a database record for each component file of the digital object and then uses the database's referential integrity enforcement to prevent one piece of the object from being deleted inadvertently. The DRS also developed a series of scripts that enforce integrity within the two-tiered architecture.

> The DRS has two main components – the management database and the file system. Each object has a database record that maintains the administrative and technical metadata for the object. The database record knows the location of the primary digital file on the disk file system. Nightly, the DRS validates the integrity of both the database and the file system by verifying that every database record the file system has a file and that every file has a database record (Harvard University Library, 2003a).

A digital repository needs to maintain the relationships between all of the components of an object.

Keeping the Context of the Object

In 1996 when the list of digital preservation goals was proposed, keeping the context of the object seemed to be a huge issue. The major concern was that organizations would have terabytes of digital files that were not usable because no one knew what they were. Over the past 10 year, the fear of having ‘orphaned’ objects has decreased. But it is still important to maintain a strong association between the intellectual metadata and the logical object. Using a well-designed preservation system should solve this issue.

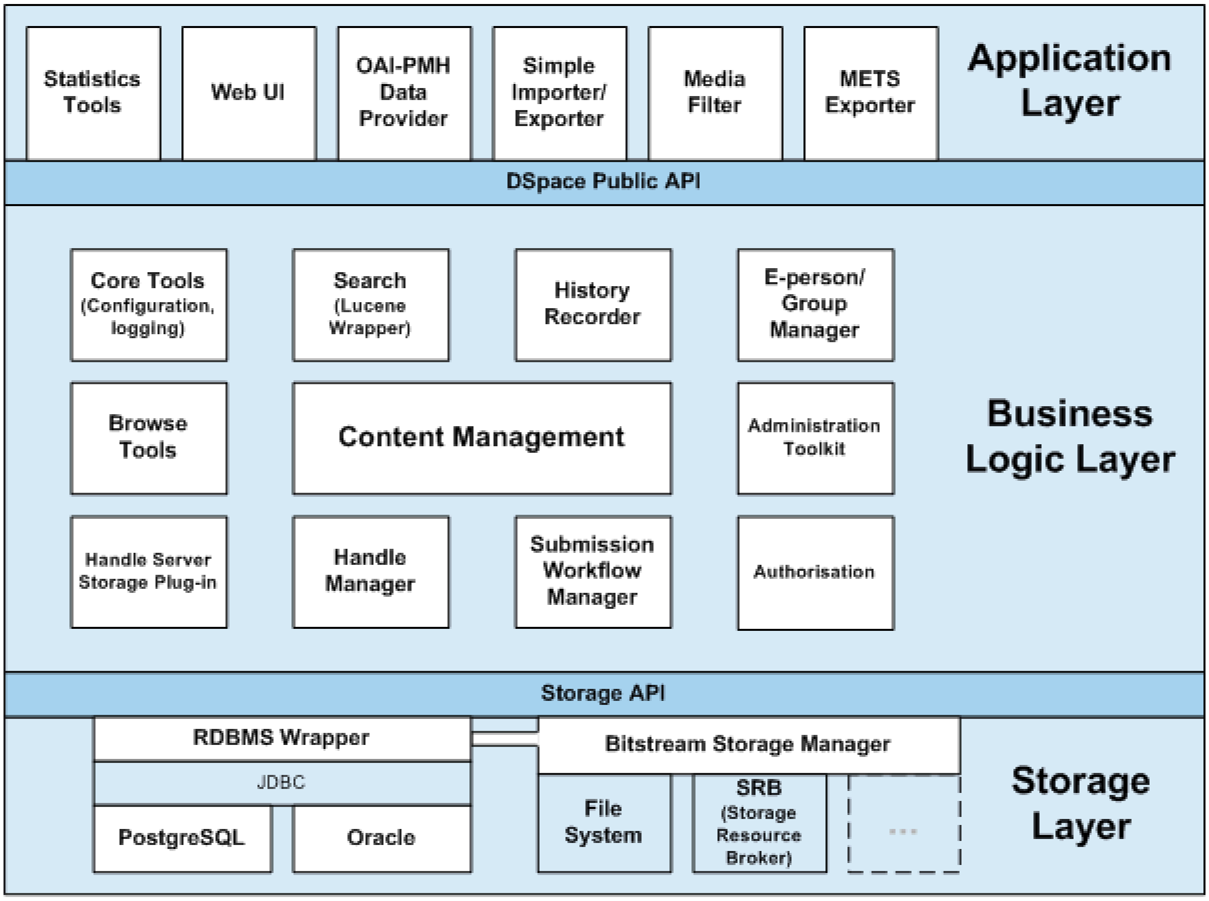
A preservation system, more commonly known as a “Digital Repository”, is a system designed as an archive for digital objects. Much of the research in digital preservation, as well as in digital libraries, has been in this area. An early attempt to find a single architecture for a universal digital repository was the Repository Architecture Protocol (RAP) – a standard access protocol for digital repositories. RAP is a modular protocol that separates the byte stream from the data type as well as the type definition from the type implementation. RAP is also extensible creating new types on demand (Blanchi & Petrone, 2001). While developed as an open source project, it is not yet clear that applications are using the RAP protocol.

Preservation System Models

Digital Repository and Institutional Repository are two terms that are often used to describe a long-term preservation system. While very similar, there are fundamental differences that need explication. A digital repository has the connotation of a back-office function – a boring, batch orientated system for collecting metadata and digital objects for general management and preservation for digital library organizations. A digital repository may have an access layer for general discovery, but it generally is thought of as a datastore. An institutional repository is both a publishing mechanism and a datastore. In opposition to a digital repository where the data generally comes from a set of trusted depositors, the institutional repository was conceived as a bottom-up, distributed data archive where data would come from the individual creators of the intellectual content – the actual authors would upload their data, create the metadata and provide an intellectual organizational structure over their own content. While different in their focus, both require "most essentially an organizational commitment to the stewardship of … digital materials, including long-term preservation where appropriate, as well as organization and access or distribution" (Lynch, 2003 p 2).

DSpace

A joint project of Massachusetts Institute of Technology (MIT) and Hewlett-Packard, DSpace is an open source repository system, which "captures, stores, indexes, preserves, and distributes digital research material" (MIT & HP, 2007). DSpace is a three-tiered system with an application layer, a business logic layer and a storage layer. The layers have a strict hierarchy and can only communicate with its immediate neighbor – the application layer can only talk to the business logic layer and may not talk with the storage layer directly. Each layer has its own API. Authentication and authorization is controlled at the application layer. Each application must insure that the people or external systems accessing the DSpace system are who they say they are. DSpace is specifically designed to allow different communities within an organization to define their "space" with a set of depositors and an intellectual organization of the contents. Access to the content is through a minimally customizable web portal (Massachusetts Institute Technology, 2004).



DSpace Overview (MIT, 2004)

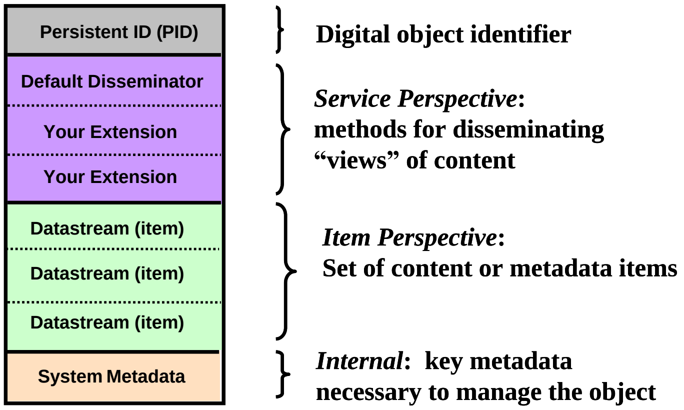
Used with permission from Massachusetts Institute of Technology

Fedora

Fedora – the Flexible and Extensible Digital Object Repository Architecture – is an open source digital repository system. Originally developed as a research project at Cornell University's Computer Science Department with funding by DARPA and NSF, it is now a joint project of the University of Virginia Library and Cornell funded by a grant from the Andrew W. Mellon Foundation (Lagoze, Payette, Shin & Wilper, 2005).

A modular design, the Fedora architecture has an extendable object model which provides a fair amount of flexibility in representing complex objects and complex relationships between objects in two perspectives – an abstract model and a function (Lagoze et al, 2005). Since the earliest designs of Fedora, a central component of the system was a "disseminator" –

the software that controlled the behaviors for metadata display and object rendering. Disseminators allow the developers to abstract the format specific constraints into layers that should allow for easier migration as the formats evolve.



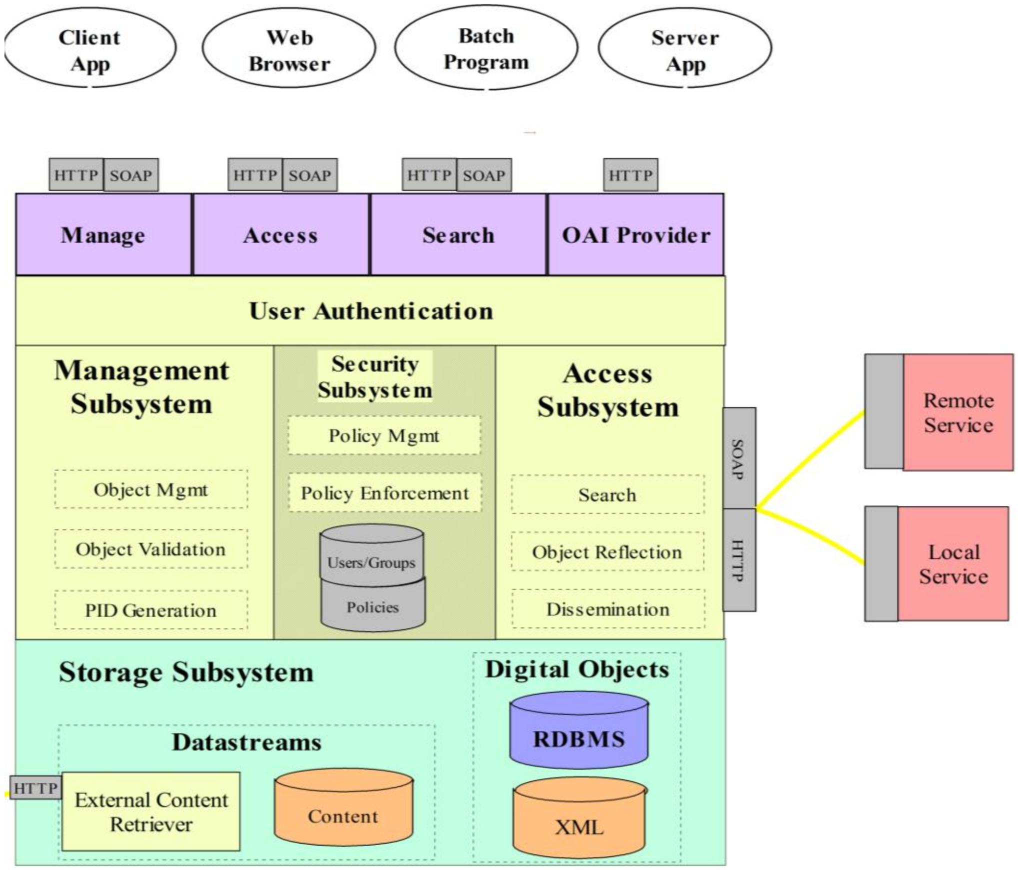
Fedora Object Model (Lagoze et al, 2005**)**

The major features of Fedora are:

- *XML submission and storage* - Digital objects are stored as XML-encoded files that conform to an extension of the Metadata Encoding and Transmission Standard (METS) schema.
- *Parameterized disseminators* - Behaviors defined for an object support user-supplied options that are handled at dissemination time.
- *Access Control and Authentication* - Although Advanced Access Control and Authentication are not scheduled until Phase II of the project, a simple form of access control has been added in Phase I of the project to provide access restrictions based on IP address. IP range restriction is supported in both the Management and Access APIs. In addition, the Management API is protected by HTTP Basic Authentication.

- *Default Disseminator* - The Default Disseminator is a built-in internal disseminator on every object that provides a system-defined behavior mechanism for disseminating the basic contents of an object.
- *Searching* - Selected system metadata fields are indexed along with the primary Dublin Core record for each object. The Fedora repository system provides a search interface for both full text and field-specific queries across these metadata fields.
- *OAI Metadata Harvesting* - The OAI Protocol for Metadata Harvesting is a standard for sharing metadata across repositories. Every Fedora digital object has a primary Dublin Core record that conforms to the schema. This metadata is accessible using the OAI Protocol for Metadata Harvesting, v2.0.
- *Batch Utility* - The Fedora repository system includes a Batch Utility as part of the Management client that enables the mass creation and loading of data objects. (Staples, Wayland & Payette, 2003)

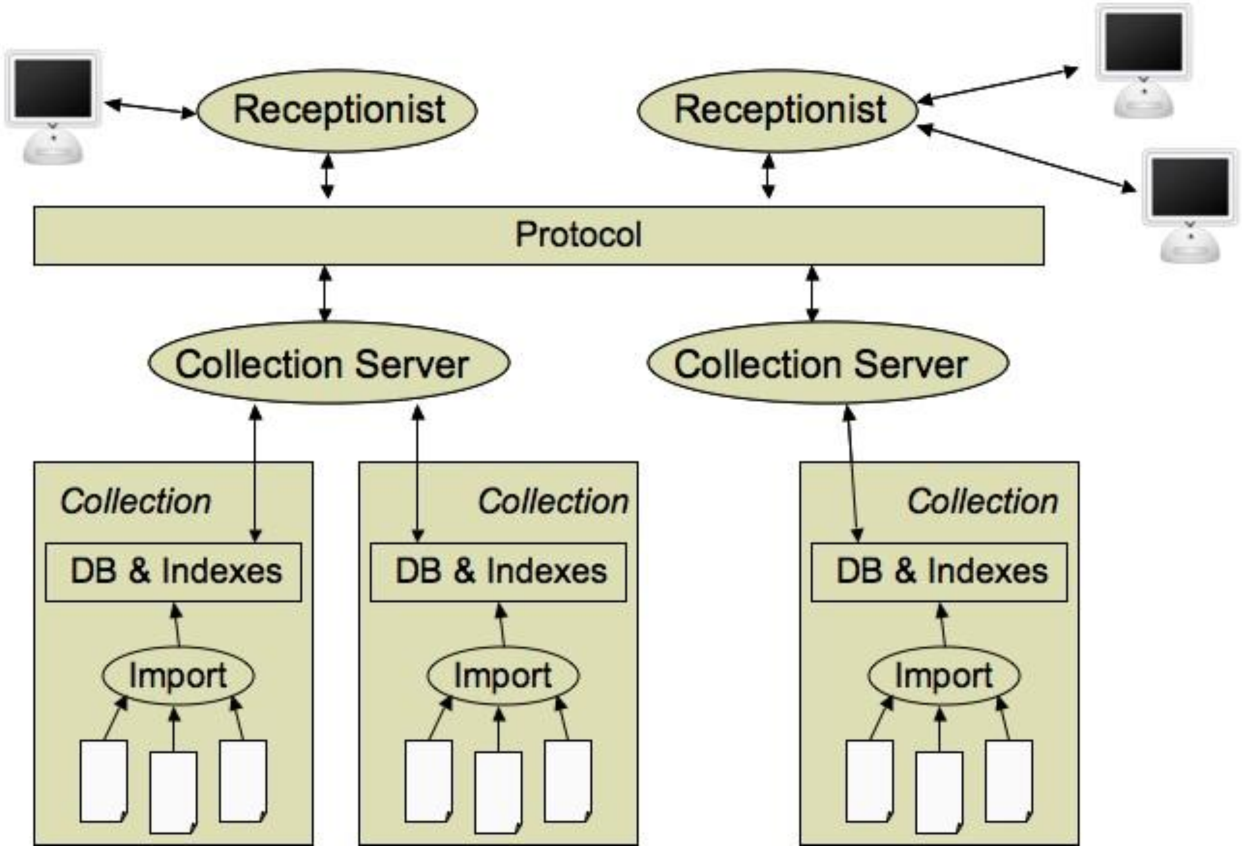
Additional functions to be added to Fedora are component management, advanced access control using an XML-oriented policy expression that can be used to enforce fine-grained object-level policies, versioning, and preservation service*s* to monitor objects, provide alerts to vulnerabilities, and perform corrective actions.



Fedora Architectural Overview (Staples et al., 2003)

Greenstone

Greenstone was developed by the University of New Zealand and first released in 2000. An open source system, it was built to be widely distributed as a means for access to digital collections. Its early tag line was "What you see – you can get!" (Witten, Bainbridge & Boddie, 2001 p 12). It has a very low technology barrier and is often used in situations where the technical infrastructure is not mature. Its early implementation was not designed for preservation. The new implementation has a service architecture. The client service manager is the "Receptionist" while the server service manager is the "MessageRouter" (Don, Bainbridge & Witten, 2002). This architecture is very modular and extendable which should allow Greenstone to mature into a preservation system at some point.

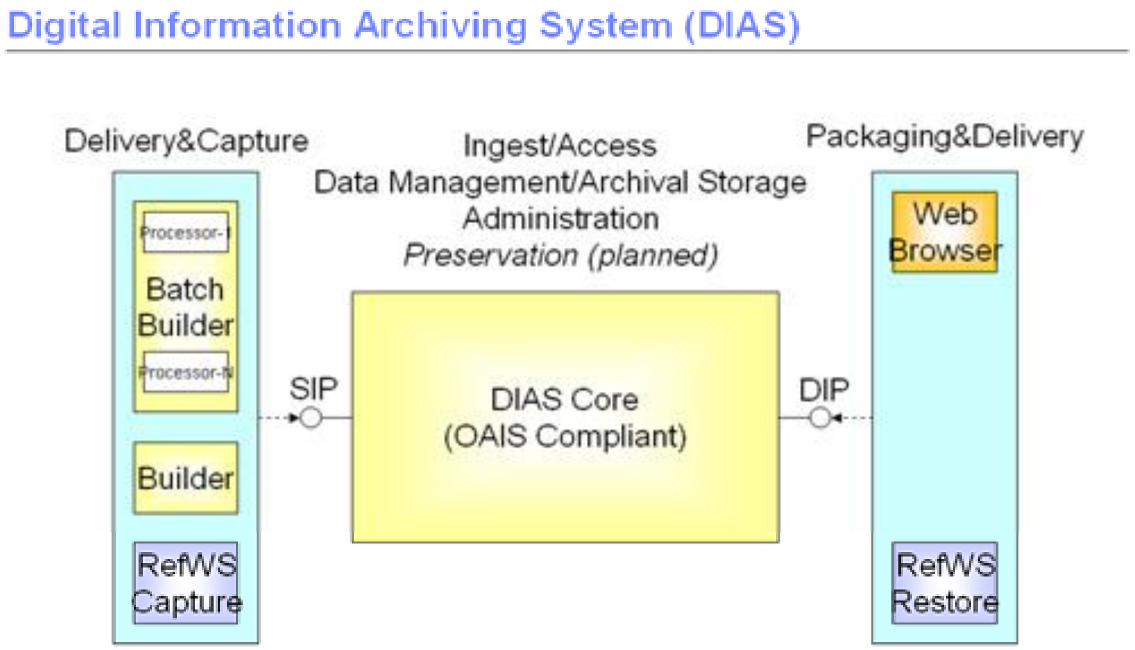


Greenstone Architecture (Olson, 2003)

Used with permission from Waikato University

Digital Information Archiving System (DIAS)

In 2000, the National Library of the Netherlands, the Koninklijke Bibliotheek (KB), began a project with IBM to create the technical infrastructure for a national digital repository for the Netherlands. Named the Digital Information Archiving System (DIAS), it is based on the OAIS preservation conceptual model (IBM, n.d.b). The system was developed using the Universal Virtual Computer (UVC) developed by Raymond Lorie, an emulation engine which allows digital objects to be reconstituted in their original form. The UVC uses a logical data scheme with type description, a format decoder, and a logical data viewer (Lorie, 2000). Rolled out in 2003, the system, known at the KB as the e-Depot, is deigned to implement the library's "Safe Place Strategy" to ensure "the transfer of digital publications form their publishing environment to a dedicated archiving environment" (Steenbakkers, 2005). The KB has taken as its primary mission to preserve the published "Scientific Record" – journal articles directly from the publisher, CD-ROMs and PDF files (Oltmans & van Wijngaarden, 2004).



DIAS Overview (IBM, n.d.a)

Digital Repository Discussion

While each of the systems describe themselves as a digital repository, they each have very different goals, organizational structures and preservation functions. DIAS and Fedora are similar in overall functional goals. Both of these systems expects to be the primary management environment for all of an organization's digital objects. They were both designed to be centrally administered, large backroom repositories with substantial batch processes for ingesting objects and their technical, structural, and intellectual metadata. DSpace has a very different overall functional goal. Its primary focus is gathering data from individuals who want to both save their files and provide public access to them. There is very little digital object management. Rather than working in a batch mode, it expects individual online real time data ingest from individuals who have some level of control over their own “space”. Greenstone has yet another primary goal – to provide access to digital collections. Like DSpace, it has minimum administrative functions, but unlike DSpace, Greenstone is not designed for individuals to upload data on their own. It is more like a web publishing application.

Although all of these repository systems were developed for universities, two, DIAS and DSpace, were developed in partnerships with commercial enterprises while Fedora and were developed solely by the universities. DIAS is a joint project of IBM and Koninklijke Bibliotheek (KB), the national library of the Netherlands, and DSpace is a project of MIT and Hewlett Packard. Even though both of these systems had commercial development partners, their business models seem to be very different. DSpace is an open source system while DIAS is not. While there do not seem to be plans to immediately turn DIAS into a commercial product, it is certainly possible. DIAS, DSpace and Greenstone all are turnkey systems. While they can be modified and customized, they are all designed for as a simple implementation. Fedora, on the other hand, is not a fully formed system, but a set of tools and modules that need to be assembled by skilled programmers, which gives it great flexibility and power but limits its applicability for many organizations.

As with their goals and organizational structures, the preservation functions of these systems are quite different. Neither DSpace nor Greenstone has any real preservation functions other than a managed datastore. Fedora and DIAS were both designed as preservation systems. While not all of the functions have yet been implemented, Fedora was designed for migration. Using standard metadata schemas, Fedora captures both technical and structural metadata to aid in future format migration. DIAS is built to provide an emulation environment so that as formats become obsolete, they will continue to function with no need to migrate. Since both DIAS and Fedora are new systems with a small install base, neither of these preservation strategies as implemented has been applied in a real production environment. While there are strong opinions on both sides of the migration-emulation divide, neither has been proven or disproved. It is probable that both will be successful up to the next computing paradigm shift when all of our previously held assumptions are shattered.

A serious look at the architectures shows very similar structures. All are n-tiered client server designs. DIAS and DSpace have three explicit layers, while Fedora and Greenstone have four. All of them separate the access logic into a layer as well as both processing logic and storage. Because of its extensive object modeling capability, Fedora has the most extendable architecture. With effort, Fedora should be able to model, store and delivery almost any type of digital object.

Repository Overview Table

	DIAS	**DSpace**	**Fedora**	**Greenstone**
Organizational Ownership	IBM and Koninklijke Bibliotheek	MIT and Hewlett Packard	Cornell University and University of Virginia	University of New Zealand
Original Functional Goal	Digital Repository	Institutional Repository	Digital Repository	Digital Content Access System
Preservation Strategy	Emulation	None	Migration	None
Preservation Metadata?	Yes	No	Yes	No
Open Source?	No	Yes	Yes	Yes

Attributes of Digital Repositories

While much has been written on the subject of Digital Repositories and many fledging systems are taking wing, little has been done to create a comprehensive list of attributes that a digital repository should include[1].

The checklist presented here is a combination of the research presented as well as the practical experience of a systems manager. The ideal use of such a checklist is at systems design or during the request for proposal (RFP) process. Unfortunately, most of us do not live in ideal circumstances. We live in organizations with legacy systems, technical environments resistant to radical change and tight resource budgets. These guidelines can be used to augment an existing infrastructure.

1. A digital repository should follow data center best practice including
    a. Regular backups
    b. Multiple copies of backups in multiple locations
    c. Backup media rotation and migration
    d. Developing a process by which files can be restored
2. To insure fixity, a digital repository should implement a checksum or digital signature on archived files and validate them on a regular schedule.
3. A digital repository should institute an annual contingency plan drill within 6 months of its initial production date.
4. A digital repository should validate digital objects when submitted for ingest. The validation should include format identification, validation and characterization.
5. A digital repository should use data formats that can be considered low risk based on the seven sustainability factors of disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, technical protection mechanisms.

[1] A draft evaluation criteria for digital repositories has been developed for a trusted external service agency called a "Certified Digital Repository" (RLG, 2005).

6. A digital repository should have processes to monitor the status of all of the file formats that the repository supports. A repository will need to determine when a change in a file format will require action for preservation.
7. A digital repository should store important technical characteristics that have been extracted from digital objects themselves.
8. A digital repository should provide provenance for all of its digital objects including a digitization date (or creation date if "born digital") and a process history.
9. A digital repository should maintain the relationships between all of the components of an object with metadata and should enforce referential integrity for the entire digital object. The system should regularly verify that all components are present and accounted for.
10. A digital repository should implement a persistent identifier service to insure long-term unambiguous access to its digital objects.

Conclusion

Digital preservation is not just good data center management. Besides keeping the bits safe, there are three additional major goals of digital preservation – keeping the files useable, keeping the integrity of the object, and keeping the context of the object (Waters & Garrett, 1996). Preserving digital information is both a technical and an organizational problem. Over the past 10 years, a great deal of progress has been made. We have begun to identify the major issues and develop solutions. We have developed a number of repositories with different models, goals and architectures. We have begun to develop community-based repositories for format information to aid preservation. We have begun to develop national strategies for preserving data. While much has been done, much more remains to be accomplished. What a wonderful opportunity for the research and the business communities to combine efforts to solve these challenges.

Future Research Directions

A report published jointly by the National Science Foundation (NSF) and the Library of Congress (LC) developed a set of research challenges for digital preservation and archiving. The four main areas of research are technical architecture for repositories, the attributes of archival digital collections, tools and technologies for digital preservation, and organizational, policy and economic issues (National Science Foundation & Library of Congress, 2003). I have expanded on a number of these issues below.

While a number of these topics have been addressed, within those broad categories, a number of substantial issues remain unexplored. Virtually all major digital preservation research initiatives have used the migration strategy. While migration will be the major strategic flow for the foreseeable future, emulation will become increasingly important especially in the gaming and entertainment environments. Emulation could also provide a more generalized solution to the thousands of multimedia publications created during the 1980s and 1990s as well as the 3D user interfaces of the systems of the future.

A major initiative needs to be undertaken to develop metrics for digital preservation. While some attempts have been made to theoretically define some metrics for risk, none have been tested in a scientifically rigorous method. We currently have no methodology of judging the quality of a preservation process. Can we measure effectiveness of a particular preservation process? How can we measure the quality of a digital object?

While format registries are an interesting start, we need much more research into understanding how to manage the multiple schema for describing both formats and metadata much less the ontologies that describe the intellectual organization of disciplines. With the rapid proliferation of formats, schemas and ontologies how will we ever find anything? We need to develop more standard preservation quality formats for some of the rapidly developing technologies – audio and moving pictures, games and interactive media.

While real, substantive work has been done in developing models of digital repositories, we have only scratched the surface. All of the repository models rightly assume other systems have created the objects. Perhaps we need to develop a theoretical and functional model for a preservation layer for generalized applications. Can we build a trusted digital object within an application prior to being sent to a repository?

Another major issue is awareness in organizations. How do we value digital assets to prioritize preservation within an organization? Can preservation become a part of information life cycles? Beyond businesses and other information producers, can digital preservation be rolled out into consumer products? Can consumer software help preserve the millions of digital photographs and personal documents that could be lost?

And of course, the major issue for digital preservation, the elephant in the room that we do not really talk about, is funding models. This is a huge issue, not just for the not-for-profit organizations that have spearheaded digital preservation research, but for businesses as well. "While the costs associated with ensuring long-term access to digital information are difficult to predict quantitatively, they are generally agreed to be significant" (National Library of Australia, n.d.a). Working with the ambiguous future makes this a difficult problem to define, much less solve. Never the less, we need to attempt to quantify total cost of ownership for a digital object.

References

Abrams, S. A. (2004). The role of format in digital preservation. *VINE, 34(2)*, 49 – 55.

Abrams, S. A. & Seaman, D. (2003). Towards a global digital format registry. *World Library and Information Congress: 69th IFLA General Conference and Council.* Retrieved on September 24, 2006 from http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Arms, C. R. & Fleischhauer, C. (2005). *Sustainability of digital formats: Planning for Library of Congress collections.* Library of Congress. Retrieved on September 24, 2006 from http://www.digitalpreservation.gov/formats/intro/intro.shtml

Bearman, D. (1999). Reality and chimeras in the preservation of electronic records. *D-Lib Magazine, 5(4).* Retrieved September 15, 2006 from http://www.dlib.org/dlib/april99/bearman/04bearman.html

Blanchi, C. & Petrone, J. (2001) *An architecture for digital object typing. Corporation for National Research Initiatives*. Retrieved September 15, 2006 from hdl.handle.net/4263537/4096

Brown, A. (2003). *Digital preservation guidance note 1: Selecting file formats for long-term preservation.* The National Archives of England, Wales and the United Kingdom. Retrieved July 10, 2006 from http://www.nationalarchives.gov.uk/preservation/advice/digital.htm

Brown, A. (2003). *Digital preservation guidance note 2: Selecting storage media for long-term preservation.* The National Archives of England, Wales and the United Kingdom. Retrieved July 10, 2006 from http://www.nationalarchives.gov.uk/preservation/advice/digital.htm

Brown, A. (2006). *The national archives digital preservation technical paper: Automatic format identification using PRONOM and DROID.* The National Archives of England, Wales and the United Kingdom. Retrieved July 10, 2006 from http://www.nationalarchives.gov.uk/preservation/advice/digital.htm

Brown, G. (2006). Virtualizing the CIC floppy disk project. *Fall Depository Library Conference, 2006.* Retrieved January 16, 2007 from http://www.access.gpo.gov/su_docs/fdlp/pubs/proceedings/06fall/brown.pdf

Buneman, P., Khanna, S., & Tan, W.C. (2000). Data provenance: Some basic issues. *Foundations of Software Technology and Theoretical Computer Science.* Retrieved April 28, 2006 from http://db.cis.upenn.edu/DL/fsttcs.abs

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS), recommendation for space data system standards, CCSDS 650.0-B-1*. Blue Book. Washington, D.C.: Author. Retrieved March 4, 2005 from http://public.ccsds.org/publications/archive/650x0b1.pdf

Corporation for National Research Initiatives. (2006). *Handle system.* Retrieved September 25, 2006 from http://www.handle.net/

Digital Curation Centre. (2005). *About the DCC*. Retrieved October 4, 2006 from http://www.dcc.ac.uk/about/

Digital Curation Centre. (2006). *Representation Information in the DCC Registry/Repository – Version 0.4.* Retrieved October 4, 2006 from http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCRegRepV04

Digital Library Federation. (n.d.). *Digital Preservation.* Retrieved on September 24, 2006 from http://www.diglib.org/preserve.htm.

Don, K., Bainbridge, D., & Witten, I. H. (2002). *The design of Greenstone 3: An agent based dynamic digital library.* Retrieved September 25, 2006 from http://www.greenstone.org/docs/greenstone3/gs3design.pdf

Fleischhauer, C. (2003). Looking at preservation from the digital library perspective. *The Moving Image, 3(2),* Fall, 96-100. Retrieved April 28, 2006 from http://muse.jhu.edu/cgi-bin/access.cgi?uri=/ journals/the_moving_image /v003/3.2fleischhauer.pdf

Foster, I., Vockler, J., Wilde, M. & Zhao, Y. (2002). *The virtual data grid: A new model and architecture for data-intensive collaboration.* Data Provenance/Derivation Workshop, October 2002. Retrieved October 10, 2006 from http://people.cs.uchicago.edu/~yongzh/papers/CIDR.VDG.submitted.pdf

Gladney, H. (2004). Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems, 22(3),* 406-436.

Harvard University Library (2003). *DRS Data Verification Process*. Internal Documentation.

Harvard University Library. (2003). *Name resolution service: Introduction and use.* (2003). Retrieved April 28, 2006 from http://hul.harvard.edu/ois/systems/nrs/nrs-intro.html

Harvard University Library. (2006). *JSTOR/Harvard Object Validation Environment.* Retrieved April 12, 2006 from http://hul.harvard.edu/jhove/

IBM. (n.d.). *Digital Information Archiving System.* Retrieved October 4, 2006 from http://www-5.ibm.com/nl/dias/index.html

IBM. (n.d.). *IBM/KB Long-term Preservation Study*. Retrieved October 4, 2006 from http://www-5.ibm.com/nl/dias/preservation2.html

International DOI Foundation. (2006). *Introductory overview: The DOI® system*. Retrieved on September 24, 2006 from http://www.doi.org/overview/sys_overview_021601.html

Kunze, J. (2003). *Towards Electronic Persistence Using ARK Identifiers.* Retrieved October 4, 2006 from http://www.cdlib.org/inside/diglib/ark/

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2005). Fedora: An Architecture for Complex Objects and their Relationships. *Journal of Digital Libraries Special Issue on Complex Objects.* 124-138. Retrieved October 5, 2006 from http://www.arxiv.org/abs/cs.DL/0501012

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. (2000*). Risk management of digital information: A file format investigation.* Council on Library and Information Resources. Washington, D.C. Retrieved on September 24, 2006 from http://www.clir.org/PUBS/reports/pub93/pub93.pdf

Library of Congress. (2006). *Metadata Encoding and Transmission Standard*. Retrieved April 28, 2006 from http://www.loc.gov/standards/mets/mets-home.html

Library of Congress. (2006). *Sustainability of digital formats: planning for library of congress collections.* Retrieved April 21, 2006 from http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml

Lorie, R.A. (2000). *Long-term archiving of digital information: IBM research report.* Retrieved on September 29, 2006 from http://domino.watson.ibm.com/library/CyberDig.nsf/7d11afdf5c7cda94852566de006b4127/be2a2b188544df2c8525690d00517082

Lynch C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL, 226.* Retrieved on September 24, 2006 from http://www.arl.org/newsltr/226/ir.html

Macey, R. (2006, August 5). One giant blunder for mankind: How NASA lost moon pictures. *The Sydney Morning Herald.* Retrieved August 8, 2006 from http://www.smh.com.au/news/national/one-giant-blunder-for-mankind-how-nasa-lost-moon-pictures/2006/08/04/1154198328978.html

Massachusetts Institute Technology. (2004). *DSpace system documentation: Architecture*. Retrieved September 25, 2006 from http://www.dspace.org/technology/system-docs/architecture.html

Massachusetts Institute Technology & Hewlett-Packard Company. (2007). Welcome to DSpace. Retrieved April 13, 2007 from http://www.dspace.org/

National Archives of Australia. (2003). *Appendix 10 – Recordkeeping cost–benefit analysis*. Retrieved September 25, 2006 from http://www.naa.gov.au/recordkeeping/dirks/dirksman/dirks_A10_cost_benefit.html

National Archives of Australia. (2003). *The DIRKS methodology – a users guide*. Retrieved September 25, 2006 from http://www.naa.gov.au/recordkeeping/dirks/dirksman/part1.html#bg3

National Archives of England, Wales and the United Kingdom. (2006). *Redaction: Guidelines for the editing of exempt information from paper and electronic documents prior to release.* Retrieved July 10, 2006 from http://www.nationalarchives.gov.uk/preservation/advice/digital.htm

National Archives of England, Wales and the United Kingdom. (2006*). The technical registry – PRONOM.* Retrieved July 10, 2006 from http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm

National Institute of Standards and Technology. (2002). Contingency Planning Guide for Information Technology Systems. Elizabeth B. Lennon (Editor). *Information Technology Laboratory*. Retrieved April 28, 2006 from http://csrc.nist.gov/publications/nistpubs/800-34/sp800-34.pdf

National Library of Australia. (n.d.) *Preserving access to digital information (PADI) initiative: Costs*. Retrieved July 15, 2006 from http://www.nla.gov.au/padi/topics/5.html

National Library of Australia. (n.d.). *Preserving access to digital information (PADI) initiative: Formats & media.* Retrieved September 25, 2006 from http://www.nla.gov.au/padi/topics/44.html

National Science Foundation & Library of Congress. (2003). It's About Time: Research Challenges in Digital Archiving and Long-term Preservation. *Final Report: Workshop on Research Challenges in Digital Archiving and Long-term Preservation*.

Ockerbloom, J.M. (2004). *Meet Fred: Format REgistry demonstration.* Retrieved October 4, 2006 from http://tom.library.upenn.edu/cgi-bin/fred?cmd=ShowDocu&&id=about

Ockerbloom, J.M. (2005). *The typed object model.* Retrieved October 4, 2006 from http://tom.library.upenn.edu/

Olson, T.A. (2003).Building collections using Greenstone. Digital Library Development Center, University of Chicago Library. Retrieved April 10, 2007 from http://www.lib.uchicago.edu/dldc/talks/2003/dlf-greenstone/

Oltmans, E. & van Wijngaarden, H. (2004). Digital preservation and permanent access: the UVC for images. *IS&T*. Retrieved October 8, 2006 from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/uvc-ist.pdf

Pearson, D. (2002). *The grid: Requirements for establishing the provenance of derived data.* Data Provenance/Derivation Workshop, October 2002. Retrieved October 10, 2006 from http://people.cs.uchicago.edu/~yongzh/papers/Provenance_Requirements.doc

Plale, B., Ramachandran, R. & Tanner, S, (2006). Data Management Support for Adaptive Analysis and Prediction of the Atmosphere in LEAD, *22nd Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology (IIPS)*, January 2006. Retrieved April 28, 2006 from http://www.cs.indiana.edu/~plale/papers/plale_IIPS06.pdf

Research Libraries Group. (2005). *An audit checklist for the certification of trusted digital repositories*. Mountain View, CA. Retrieved on October 6, 2006 from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

RLG & OCLC. (2002). *Trusted digital repositories: Attributes and responsibilities: An RLG-OCLC report.* Retrieved April 28, 2006 from http://www.rlg.org/longterm/repositories.pdf

Rothenberg, J. (1999). *Avoiding technological quicksand: Finding a viable technical foundation for digital preservation.* The Council on Library and Information Resources. Retrieved July 15, 2006 from http://www.clir.org/PUBS/reports/rothenberg/contents.html

Rothenberg, J. (1999). *Ensuring the longevity of digital information*. The Council on Library and Information Resources. Retrieved July 15, 2006 from http://www.clir.org/pubs/resources/articles.html

Saltz, J. (2002). *Data provenance.* Data Provenance/Derivation Workshop, October 2002. Retrieved October 10, 2006 from http://people.cs.uchicago.edu/%7Eyongzh/papers/ProvenanceJS10-02.doc

Simmhan, Y., Plale, B., & Gannon, D. (2006). A Performance Evaluation of the Karma Provenance Framework for Scientific Workflows, *to appear IPAW'06*. Retrieved April 28, 2006 from http://www.cs.indiana.edu/~plale/papers/SimmhanIPAW06.pdf

Smith, A. (2006). Distributed preservation in a national context: NDIIPP at mid-point. *D-Lib Magazine, 12(6).* Retrieved October 11, 2006 from http://www.dlib.org/dlib/june06/smith/06smith.html

Stanescu, A. (2004). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *D-Lib Magazine, 10(11).* Retrieved July 15, 2006 from http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

Staples, T., Wayland, R., & Payette, S. (2003). The Fedora project: An open-source digital object repository management system. *D-Lib Magazine*, 9(4). Retrieved October 5, 2006 from http://www.dlib.org/dlib/april03/staples/04staples.html

Steenbakkers, J. F. (2005). Digital archiving in the twenty-first century: Practice at the national library of the Netherlands. *Library Trends, 54(1),* 33 – 56.

Technical Advisory Service for Images (TASI). (2006). Choosing a file format. Retrieved October 11, 2006 from http://www.tasi.ac.uk/advice/creating/format.html#fo3

UKOLN. (2006). *Good practice guide for developers of cultural heritage web services.* Retrieved July 15, 2006 from http://www.ukoln.ac.uk/interop-focus/gpg/

Waters, D., & Garrett, J. Eds. (1996). *Preserving digital information: Report of the task force on archiving of digital information.* Washington, D.C. and Mountain View, CA: The Commission on Preservation and Access and the Research Libraries Group. Retrieved March 4, 2005 from http://www.rlg.org/ArchTF/

WC3. (2001). URIs, URLs, and URNs: Clarifications and Recommendations 1.0. *Report from the joint W3C/IETF URI Planning Interest Group, W3C Note 21, September 2001.* Retrieved October 6, 2006 from http://www.w3.org/TR/uri-clarification/

Witten, I.H., Bainbridge, D., & Boddie, S.J. (2001). Greenstone: Open-source digital library software. *D-Lib Magazine, 7(10).* Retrieved October 5, 2006 from http://www.dlib.org/dlib/october01/witten/10witten.html

Weibel, S., Jul, E., & Shafer, K. (1995). *PURLs: Persistent uniform resource locators*. Retrieved October 8, 2006 from http://purl.oclc.org/

Additional Readings

Arms, W. (1999). Preservation of scientific serials: Three current examples. *Journal of Electronic Publishing, 5*. Available at http://www.press.umich.edu/jep/05-02/arms.html

Atkins, D. E. (2003). Revolutionizing science and engineering through cyber-infrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure. *Directorate for Computer & Information Science & Engineering.* Available at http://www.cise.nsf.gov/sci/reports/atkins.pdf

Barnett, B., Bishoff, L., Borgman, C., Caplan, P., Hamma, K. & Lynch, C. (2003). Report of the Workshop on opportunities for research on the creation, management, preservation and use of digital content. Available at http://www.imls.gov/pubs/pdf/digitalopp.pdf

Bearman, D. (1999). Reality and chimeras in the preservation of electronic records. *D-Lib Magazine 5*(4). Available at http://www.dlib.org/dlib/april99/bearman/04bearman.html

Cantara, L. (2003). Introduction. In: L. Cantara, (Ed.) *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation.* Available at http://www.diglib.org/preserve/ejp.htm

Carlin, J. W. (2004). ERA: An archives of the future for the future. *Prologue, a quarterly publication of the National Archives and Records Administration 36 (1),* Spring 2004. Available at http://www.archives.gov/publications/prologue/2004/spring/archivist.html

Cathro, W. (2004). Preserving the outputs of research. *Archiving Web Resources Conference, Canberra, Australia, November 9-11 2004*. Available at http://www.nla.gov.au/nla/staffpaper/2004/cathro1.html

Crow, Raym. (2002). *SPARC institutional repository checklist & resource guide.* The Scholarly Publishing & Academic Resources Coalition, American Research Libraries. Available at www.arl.org/sparc

Davis, S. (2002). Digital preservation strategy. The National Archives of Australia Agency to Researcher Digital Preservation Project. Available at http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf

Falk, Howard (2003). Digital archive developments. *The Electronic Library*, *21*, 375–379. Available at http://caliban.emeraldinsight.com

Flecker, D. (2001). Preserving scholarly e-journals. *D-Lib Magazine*, *7*(9). Available at http://www.dlib.org/dlib/september01/flecker/09flecker.html

Gadd, E., Oppenheim, C., & Probets, S. (2003). The Intellectual property rights issues facing self-archiving – key findings of the RoMEO project. *D-Lib Magazine*, *9*(9). Available at http://www.dlib.org/dlib/september03/gadd/09gadd.html

Gilliland-Swetland, A., Eppard, P. (2000). Preserving the Authenticity of Contingent Digital Objects. *D-Lib Magazine, 6(7/8).* Available at http://www.dlib.org/dlib/july00/eppart/07eppard.html

Granger, S. (2002). Digital preservation and deep infrastructure. *D-Lib Magazine 8*(2). Available at http://www.dlib.org/dlib/february02/granger/02granger.html

Harnad, S. (2001). The self-archiving initiative: Freeing the refereed research literature online. *Nature 410* (April 26), 1024-1025. Available at http://www.ecs.soton.ac.uk/~harnad/Tp/nature4.htm

Hart, P.E. & Liu, Z. (2003). Trust in the preservation of digital information. *Communications of the ACM, 46(6),* 93-97. Available at http://doi.acm.org/10.1145/777313.777319

Harvard University Library. (2003) Report on the planning year grant for the design of an e-journal archive. In: L. Cantara, (Ed.) *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation.* Available at http://www.diglib.org/preserve/ejp.htm

Hedstrom, M. & Ross, S. (2003). Invest to save: Report and recommendations of the NSF-DELOS working group on digital archiving and preservation. Available at http://eprints.erpanet.org/archive/00000048/01/Digitalarchiving.pdf

Hunt, A. & Thomas, D. (2002). Software archaeology. *IEEE Software, 19(2),* March/April 2002.

Inera, Inc. (2001). *E-journal archive DTD feasibility study.* Prepared for the Harvard University Library, Office of Information Systems, E-Journal Archiving Project. Available at http://www.diglib.org/preserve/hadtdfs.pdf.

Kaplan, E. (2002). Response to "preserving software: why and how". *Iterations: An Interdisciplinary Journal of Software History, 1(13),* September 2002, 1-3. Available at http://www.cbi.umn.edu/iterations/kaplan.html

Kling, R., Spector, L. B., & Fortuna, J. (2003). The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central. *Journal of the American Society for Information Science and Technology*, *55*(2), 127–148. Available at http://www3.interscience.wiley.com/

Library of Congress. (2003). Preserving our digital heritage: Plan for the national digital information infrastructure and preservation program: a collaborate initiative of the Library of Congress. Available at http://www.digitalpreservation.gov/about/planning.html

Lord, P. & Macdonald, A. (2003). E-Science curation report : Data curation for e-science in the UK : an audit to establish requirements for future curation and provision. *The JISC Committee for the Support of Research.* Available at http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Liu, Z. (2003). Trends in transforming scholarly communication and their implications. *Information Processing & Management*, *39*, 889-898. Available at http://www.elsevier.com/

Maniatis, P., Rosenthal, D., Roussopoulos, M., Baker, M., Giuli, T. J., & Muliadi, Y. (2003). Preserving peer replicas by rate-limited sampled voting. *ACM Symposium on Operating Systems Principles archive – Proceedings of the nineteenth ACM symposium on Operating systems principles table of contents*. Available at http://portal.acm.org/

Marcum, D.B. (2003). Research questions for the digital era library. *Library Trends, 51(4),* Spring 2003, 636-651.

National Science Foundation Office of Cyberinfrastructure. (2005). NSFs cyberinfrastructure vision of 21st century discovery. Available at http://www.nsf.gov/od/oci/CI-v40.pdf

Pearson, D. (2001). Medical history for tomorrow – preserving the record of today. *Health Information and Libraries Journal*, 18. Available at www.blackwell-synergy.com/www.blackwell-synergy.com/

Pinfield, S., & James, H. (2003). The digital preservation of e-prints. *D-Lib Magazine, 9*(9*).* Available at http://www.dlib.org/dlib/september03/pinfield/09pinfield.html

Reich, V., & Rosenthal, D. (2004) Preserving today's scientific record for tomorrow. *BMJ: British Medical Journal*, *328*(7431). Available at http://www.bmjjournals.com

Rosenthal D., Lipkis T., Robertson T., & Morabito S. (2005). Transparent format migration of preserved web content. *D-Lib Magazine, 11*(1). Available at http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Smith, B. (2002). Preserving tomorrow's memory: Preserving digital content for future generations. *International Preservation News, 29*, May 2003, 4-9. Available at http://www.ifla.org/VI/4/news/ipnn29.pdf

Spedding, V. (2003). Data preservation : Great data but will it last? *Resource Information 5*, Spring 2003**.** Available at http://www.researchinformation.info/rispring03data.html

Waters, D. (2002). Good archives make good scholars: Reflections on recent steps toward the archiving of digital information. Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub107/waters.html

Waters, D. (2006) Preserving the Knowledge Commons. In *Knowledge as a Commons: From Theory to Practice*, Elinor Ostrom and Charlotte Hess, eds. *Understanding* Cambridge: MIT Press, forthcoming.

Wheatley, P. (2004). Institutional repositories in the context of digital preservation. *DPC Technology Watch Series Report, 04-02*. Available at http://www.dpconline.org/docs/DPCTWf4word.pdf

Woodyard, D. (2004). Significant property: Digital preservation at the British Library. *VINE, 34(1)*, 17-20. Available at http://ninetta.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=862526

Zabolitzky, J.G. (2002). Preserving software: Why and how. *An Interdisciplinary Journal of Software History, 1 (13),* September 2002, 1-8. Available at http://www.cbi.umn.edu/iterations/zabolitzky.html