Date : 01/07/2008

Ingest Strategies for Digital Libraries: the Challenges of Handling Portable Objects

Adam Rusbridge

and

Seamus Ross
{a.rusbridge, s.ross}@hatii.arts.gla.ac.uk
Humanities Advanced Technology and Information Institute
University of Glasgow

Meeting: **84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**

Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Abstract

Increasingly we have come to think of institutional repositories as the primary storage and access locations for the deluge of digital materials. However, we continue to employ portable physical carriers for the storage and transfer of digital information, from portable hard drives, USB and Flash drives, through to DVD-ROMs, CD-ROMs and floppy disks. As concerns about the fragility and obsolescence of physical carriers continues to mount, it is likely that digital material will increasingly be transferred back to digital repository systems where curation of these objects will be simplified. The department of Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow ran the Packaged Object Ingest Project (POIP) to investigate the requirements, procedures and challenges that arise from ingesting digital objects on portable media into a repository. The Packaged Object Ingest Project provides a first step towards defining best practice for ingest from portable carriers. The work that lies behind this paper was carried out under ERPANET, a project funded by the EU under the FPS programme.

The challenges of ingest from portable carriers should not be underestimated. A range of material was selected, designed to provide a realistic testbed. We studied a variety of published and unpublished CD-ROMs, published and unpublished 3.5" floppy disks, and unpublished 5.25" disks. We found the content and variations in production often posed challenging intellectual issues. How should we catalogue and store objects containing complex multimedia software applications or those containing seemingly unrelated collections of documents? Particularly with floppy disks, we encountered examples where the collections spanned multiple disks. What is the most appropriate way to catalogue these objects and retain the collection's provenance? The increased reliance on networked storage and delivery along with the rate of technological change means hardware obsolescence is a tangible risk. We wished to investigate which issues associated with ingest were unique to physical carriers as opposed to network delivered content.

Our approach was to develop a small testbed database and repository, focusing on the functions related to ingest of the contents of portable objects. We selected what we then perceived to be the most suitable preservation metadata schema, the National Library of New Zealand Preservation Metadata Schema, as our underlying data structure. Having built a simple interface, we ingested the contents represented on a range portable objects, paying particular attention to address issues related to the variety of content and production values.

Unsurprisingly, we discovered that the manual ingest of digital contents is a slow and expensive process, particularly for complex objects represented in many files with varying levels of binding. The array of file formats encountered and range of appraisal decisions meant that any ingest operator would be required to have not only a high level of technical skill, but often familiarity with the subject domain and an appreciation of the relevance of the contents. It was necessary to render the object to appraise it's contents. Attempting to render the digital objects on different hardware and operating systems resulted in either failure, partially broken or severely limited renderings on several occasions. The range of file formats found on most portable objects meant that significant preservation metadata was required and though extraction and creation was expensive at first, the resource cost decreased with repetition.

This paper describes the research that we conducted, summarises our findings, and presents the practical guidance that we developed as a result. In conclusion we indicate areas where further research might be appropriate to refine our conclusions and to address questions where our results were less than conclusive.

Introduction[1]

Increasingly we have come to think of institutional repositories as the primary storage locations for the deluge of digital materials. However, we continue to employ portable physical carriers for the storage and transfer of digital information, from portable hard drives, USB and Flash drives, through to DVD-ROMs, CD-ROMs and floppy disks. Physical carriers may be in use for a variety of reasons. For example, individuals and organisations may establish a self-archiving policy, storage capacity may demand the transfer of digital content to a physical carrier, or an object may be most effectively distributed via physical means. However, as concerns about the fragility and obsolescence of physical carriers continues to mount, it is likely that digital material will increasingly be transferred back to digital repository systems where curation of these objects will be simplified.

The challenges of ingest from portable carriers should not be underestimated and we were concerned that this area is not yet thoroughly understood. The department of HATII within the University of Glasgow ran the Packaged Object Ingest Project (POIP) to investigate the requirements, procedures and challenges that arise from ingesting digital objects on portable media into a repository. The Packaged Object Ingest Project provides a first step towards defining best practice for ingest from portable carriers.

Our objective was to explore the challenges that arise from the variations in content and production found on published and unpublished physical carriers, such as CD-ROMs, 3.5" floppy disks and 5.25" disks. Although our study did not include DVDs and USB memory sticks, the challenges these present are analogous to those we investigated. We hoped the experiment would allow us to better understand the skills and effort required and cost involved in metadata generation, and in particular to formalise workflows and identify processes where automation may be possible.

Previous work

The National Library of New Zealand has been careful to work on developing a strong intellectual framework to form the basis of a mature digital library system. A recent review of their Digital Library Development work [1] highlighted the need to understand the methods used when manipulating small digital objects. Further conclusions showed the need for related procedures for handling and ingesting packaged, physical objects.

In related work, the CEDARS Project produced a test synthesis of their implemented repository [2], the National Archives have been developing the CAMS database [3], the San Diego Supercomputer Center ingested one million Usenet records into a database [4], the OCLC/RLG PREMIS group are further developing the OCLC/RLG Preservation Metadata framework [5] through practical research. The Amsterdam Municipal Records Office investigated some of the issues associated with the transfer, identification and archival of nine CD-ROMs [6], the National Library of Australia has been actively experimenting with repositories and the National Library of New Zealand has begun a practical implementation and investigation of their theoretical work.

Our Approach

The Packaged Object Ingest Project was designed to provide a realistic testbed upon which the workflow of ingest could be explored[2]. Readers familiar with the Open Archival Information System (OAIS) model [7] will notice that our study resides at the juncture between object creation and preparation of Submission

1 The Authors wish to thank (a) Steve Knight from the NLNZ for providing us with access to recent work on the NLNZ preservation metadata standard, the NLNZ technical metadata extraction tool and his expertise. (b) Foetini Avarani, former lecturer in HATII, for her guidance with defining the minimal description dataset we should use for the prototype. (c) Andrew McHugh, HATII Resource Development officer, for contributing to the development of the underlying database.

2 The research for the Packaged Object Ingest Project was undertaken in 2003. While technology and standards may now provide solutions to some of the challenges presented here, this paper is more concerned with the underlying concepts that should be considered when considering ingest of information held on physical carriers, and establishing and developing digital repository infrastructures.

Information Packages (SIPs). We anticipated a study that identified procedures to transfer information from physical carriers to a digital repository and an investigation of bibliographic and preservation metadata needed to support this. We wished to explore issues of media vulnerability, along with an investigation into how physical carriers had been used by different groups to store data. What metadata must we generate to understand the information content and the context in which it existed? With such a spectrum of physical carriers and the variety of formal and informal production techniques, will we be able to identify processes where automation could be used to lower cost?

Although we designed the project to to focus solely on ingest, it became apparent that many of the issues identified were associated with the selection, appraisal and long term curation of data. We have highlighted these issues, although were not able to explore these thoroughly within the context of the POIP project. HATII is actively taking these related issues forward through participation in the DAFD programme[3] to audit awareness, policies and practice for data curation and preservation, and the Digital Curation Centre[4] to support expertise and practice in data curation and digital preservation.

Our approach was to develop a lightweight prototype repository focusing on the functions related to the ingest of digital information. We developed our repository using a MySQL (version 4.0.11) database and PHP (version 4.3.1) user interface all operating on a Linux-based server (Mandrake 9.1). By writing a simulation database we were given flexibility to experiment with different workflows, retaining control over the particular processes we wished to focus on. Our experiment pre-dated the PREMIS metadata schema [12], and so we added fields for technical metadata corresponding to what we then perceived to be the most suitable preservation metadata schema, the National Library of New Zealand Preservation Metadata schema [8], and fields for bibliographic and descriptive metadata corresponding to the MARC21 standard [9]. Having built a simple interface, we ingested the information contents stored on a range of physical carriers, identifying and recording the issues that arose from the variety of information content and methods of production.

We determined that initially a simple, linear ingest workflow would best expose the challenges of ingest and allow us then to identify a streamlined procedure. A digital object is uploaded to the repository where it is held on a secure filesystem. Objects were uploaded via a HTTP form however a more complete implementation would likely utilise transfer mechanisms such as Secure Copy (SCP) or a secure version of the File Transfer Protocol (FTP). Technical metadata was created for each file. We expected that some complex digital objects would be composed of several files of identical type, each requiring similar technical metadata, and so added functionality to allow the re-population of fields from previous records. Bibliographic metadata was created for the object as a whole and to maintain an accurate provenance record we recorded the processes applied.

Sample Dataset

Forty five physical carriers were selected at random to produce a stratified sample for ingest into our repository. Achieving a truly random sample is difficult; for our test dataset we turned to the personal data archive of one of the authors of this paper and selected approximately twenty percent of the physical carriers found within. Details of the physical carriers included in our pilot study can be found in Appendix 1. Woodyard [10] has highlighted many of the ingest complexities associated with floppy disks, such as the number of disks per object, the OS required (eg. Mac vs Windows), the hardware required (High Density Floppy vs. Variable Speed Floppy) and the formatting of the disk (eg. FAT32 vs. UFS). Similar technical concerns apply to CD-ROMs. Although the physical carriers in our dataset were chosen at random, we feel the spread of content provides a reasonably representative sample of physical carriers and content a digital library may encounter, and that the issues encountered may be analogous to those found on other carriers.

Overview of Physical Carriers

Four types of physical carriers were included in our study: 5.25" floppies, 3.5" floppies, CD-ROMs and CD-Rs. While we may classify material by media type it is perhaps more useful to consider the content and

3 http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframework.aspx
4 http://www.dcc.ac.uk

production values. Below we identify the categories into which our sample set may be sorted, although in some cases the boundaries between classifications may be somewhat blurred.

Published Static Content CD-Roms
In our test set, these were generally found to contain clear, hierarchical directory structures of published documents. Where a user interface was provided it was implemented in simple HTML. The purpose of the user interface was to aid identification and resource discovery.

Multimedia CD-Roms
These objects specified minimum hardware requirements and often required specific software environments for execution. In some cases this required software was also found on the disk. These objects tended to provide the user with an immersive and interactive multimedia environment where the experience felt just as important as the information content. Ensuring access to the object may mean more than just ingesting the contents of the disk and preservation of associated software and hardware may be necessary.

Unpublished CD-Rs
These tended to be backups or copies of work made by authors or creators for personal use or self-archiving. Frequently the content was poorly documented, if documented at all. Although the information objects found on the physical carrier may have been related in some context (for example, containing a personal archive of content) this was not immediately evident and the objects appeared without any apparent association.

Published 3.5" Floppy Disks
Before CDs became a popular mode of dissemination floppies were predominately used. Limitations of capacity and sophistication of technology means we largely found these contained directory structures of documents. The low capacity means aggregated collections of publications often span several disks.

Unpublished 5.25" and 3.5" Floppy Disks
The portability and storage opportunities of floppy disks led to their use as a medium for data transfer and personal backup. The casual usage of these disks meant they were often poorly documented, both digitally and physically.

Workflow of Ingest

The paragraphs below describe the stages and procedures we followed during ingest of the information content on our sample dataset into our prototype repository. We acknowledge that although we have attempted to identify distinct procedures this is somewhat difficult as many are inter-related companion stages to a larger process. In OAIS terminology, we reiterate that these are component processes associated with the preparation of Submission Information Packages (SIPs).

Selection
The physical carriers selected for this study were chosen from several hundred examples pulled from the shelves, boxes and desk drawers of one of the authors. We attempted to make the selection as random as possible but we did wish to ensure that some CD-ROMs, CD-Rs and floppies were included in published and unpublished form. We note that our approach has led to a somewhat artificial ingest environment, however our random selection was intended to provide a cross section of the packaged objects we feel digital libraries are likely to encounter. A working digital repository should in practise be ingesting material according to an explicit selection and appraisal policy. Such a policy will clarify and resolve many of the relevance and context issues we encountered during cataloguing and metadata creation. Likewise, issues of registration and ingest preparation may be alleviated with explicit selection and appraisal guidance.

Registration
For each of the forty-five items we attempted to contact the originator of the work, be it the author or publisher. Where an organisation obtains material via legal deposit such contact may be unnecessary but copyright and licensing may render this compulsory, especially where subsequent access is provided. Obtaining permission to use material in this experiment proved challenging. For example, one CD-ROM contained the proceedings of an Institute of Electrical and Electronics Engineers (IEEE) conference. The IEEE were concerned that we were redistributing their work, but agreed to inclusion following clarifications

regarding our objectives and the terms of distribution. It is clear that appropriate collection management and access policies and procedures will need to be followed.

Quarantine and Virus Checking
The National Archives of Australia have proposed a quarantine period [11] of between four and eight weeks during which an object should not be accessed or executed. This will allow for new software viruses to be detected and ensure virus checking software has been updated. Following this period the digital object should be scanned (we used the Sophos Anti-Virus software[5] along with the then most recent virus identity files (IDEs)). At the time of our experiment we did not check for the presence of spyware and bots, although the recent growth in their use means this is now a necessary procedure. None of the disks in this study had been accessed in the last two years so we felt that implementing the quarantine step was unnecessary.

Ingest Preparation
If a physical carrier is sealed within its original packaging, its museological value as a rare artefact may far outweigh the value of its information content. Before inspecting the content of a physical carrier, the packaging was examined. Published materials often provided clear descriptions of the information content held on the associated physical carriers, along with any hardware and software system requirements. “Entry” or “start” points to objects consisting of more than one file could also generally be established from the documentation. However, many physical carriers were not labelled in a useful way – this was especially true for unpublished works - and it was necessary to install, execute and view the objects to determine their entry points, technical properties and bibliographic information. This process was also necessary to confirm the object's completeness and to understand the functionality of the digital content.

Many objects were not cross-platform compatible and so required inspection or execution under the original operating system. This requirement manifested itself in unexpected ways. For example, one CD-ROM containing HTML pages did not work. Upon closer inspection we discovered this was a result of differences in how Windows and Linux systems handle case sensitivity of characters in URLs. The digital content had been produced under Windows, without case sensitivity, and so some URL references contained capitalisation that did not correspond to the actual file name. Viewing this object under Linux failed, where differences in capitalisation are meaningful. In our test environment, viewing content and extracting bibliographic and technical metadata was a demanding task. Ingest technicians may need familiarity with a wide range of utilities and platforms to support the process of inspection and metadata generation. Automated metadata extraction tools or limiting the variety of material that a single repository could ingest may simplify the process.

Verification
Following transfer to the repository a hash-based verification method should compare the original and 'preservation master' to confirm that no errors have been introduced during transfer and the object's integrity remains intact. Verification should not just confirm that individual files remain intact, the completeness of the objects should be confirmed either through inspection or by execution. To perform this step a dedicated non-critical test environment should be used, where a copy of the program can be executed or object inspected from a non-privileged user account with file permissions set to read and execute only. It is advisable to take these measures to prevent malicious scripts from harming the host system, but equally important to prevent the object from inadvertently being altered.

Description and Cataloguing
Appropriate bibliographic and technical metadata should be created to support both resource identification and discovery, and in support of ongoing digital curation. The challenges that result from identifying and generating appropriate metadata are described extensively throughout the remainder of this report.

Archiving
Digital objects should be transferred to an appropriate and secure permanent storage location where their metadata allows later identification. Measures should be taken to ensure security and integrity during transfer. It may be necessary to encapsulate complex objects in a single archiving file such as that produced

5 Primarily because Sophos was the University of Glasgow approved anti-virus software.

by the GNU ‘tar’ package. Regular backup and medium refreshment should occur and the integrity of digital objects should periodically be verified. Such activities are outside the scope of this paper.

Five Case Studies

Twenty of the forty five digital objects in our sample dataset have been ingested into the testbed repository. The twenty objects ingested were chosen from the dataset as they consisted of the majority of physical carriers and contained a wide variety of content and production values, allowing us to identify a broad set of issues within the limitations of our available resources. Although we acquired a 5.25” floppy drive the electronics had failed and time constraints meant we were unable to further investigate the content of such media appropriately. Below we have selected five scenarios which together summarise many challenges faced and questions asked.

Ingest of ‘The Digital Culture’ Floppy Disk[6]

This physical carrier was a floppy disk labelled with the handwritten text “The Digital Culture: Maximising the Nations Investment, Word 6.0 files". No other packaging or documentation was provided with the disk. Upon investigation, the disk was found to contain three folders with a total of twenty seven files in Rich Text Format (RTF), Microsoft Word and Portable Document Format (PDF). Upon initial inspection, there did not appear to be any meaningful relationship between individual files and folders. For example, one PDF file contains an application for the renewal of a Canadian Drivers Licence. One of the folders contains several document files purporting to be ‘biographical sketches’. Another folder holds a report on digital archaeology split by chapter across several files. One file on the disk was a Rich Text Format (RTF) document entitled ‘tamaro.rtf’. Viewing the contents of this file a reader is presented with the header “PRIVATE AND CONFIDENTIAL”. Access controls for each object ingested ought to be specified depending on the contents of the file and purpose of the repository.

Although these files were found on the same disk, should they be considered part of the same discrete object? The individual files must be inspected to pick out relationships and where files are catalogued as distinct and individual objects, the context in which they were discovered should be recorded. A relationship between such individual files may later be identified. Determining these relationships between objects should be addressed by staff conforming to a selection and appraisal policy. Content is presumably being ingested for some purpose and this should mean collection decisions – the selection of, determining relationships between, and discarding of material – all ought to be guided by context and background.

Ingest of ‘A^{2}PAW’ CD-R[7]

The Artist’s Archive, Preservation & Access Workbench (A^{2}PAW) CD-R contains a backup of the A^{2}PAW website, associated documentation, the OpenOffice installation package and a README file. The backup of the website consists of a collection of PHP and HTML files. To correctly execute these files and view their information contents, a web server application running PHP is required. It is not clear that a community organisation is currently preserving such software on behalf of individual repositories. To what extent should repositories be responsible for the preservation of software? Repository administrators may need to establish policies to determine whether they should locally take measures to ensure this software is also preserved, or if it is only essential to specify the additional software required. Attempting to view the website, we discovered the associated and required MySQL database was not stored on the CD-ROM. In its current form the object is incomplete and cannot be rendered correctly. To avoid this scenario steps should be taken to verify an object's completeness at delivery and ingest.

Ingest of ‘Clarysse Mac’ Floppy Disk[8]

The “Clarysse Mac” floppy disk contained a speech written as a text document. The physical carrier

---

6 The authors would like to acknowledge Seamus Ross, ‘The Digital Culture: Maximising the Nations Investment, Word 6.0 files’

7 The authors would like to acknowledge the Humanities Arts and Technology Information Institute, ‘APAW’, ©2001 HATII

8 The authors would like to acknowledge Seamus Ross, ‘Clarysse Mac’, ©1999

presented us with challenges of media obsolescence. Macintosh File System (MFS) is a disk file system introduced in 1984 by Apple Computer for storing files on 400K floppy disks. MFS was termed a flat file system because it simulated rather than supported a hierarchical directory structure. MFS was never supported by Windows and UNIX and since MacOS 8.5 the format was no longer supported by Apple. Apple developed a variable speed floppy drive for use with the MFS which enabled the floppy disks to store 400KB instead of the industry standard 360KB. Similarly, high density 720KB floppies were able to store 800KB. To achieve this the drive span the disk at different speeds depending on the distance from the hub. At the time of our experiment the hardware drives produced by Apple were the only devices that could access disks formatted according to the MFS file system. Although a drive may have had the physical capability to read the disks, upgrading beyond MacOS 8.5 released in 1998 would have disabled the machines ability to read the content.

We were unable to access the contents of the 'Clarysse Mac' floppy disk on either Linux or Windows based machines. We inspected the disk using a UNIX utility compatible with Hierarchical File Systems, Apple Computer's disk file system that followed MFS. When these utilities failed to recognise the contents of the disk we determined we had an MFS formatted floppy. After gaining access to an old working Apple Macintosh in the Glasgow University Library Computing Support Centre we discovered the disk contained a Word for Macintosh 5.1 file named ‘speech_tomo’ along with several fonts, all named with the prefix ‘supergreek’.

Copying the contents of the original floppy to a 1.44MB Windows-formatted floppy disk allowed us to ingest the material into our repository. Microsoft Word 2000 was able to successfully render the file contents and so we archived the file in its original format. However, although the Word document had successfully copied, the associated font files failed to copy correctly from the original floppy. Although this appeared not to affect the contents of the document, the appearance may have been altered and we could not be sure that loss had not occurred.

Ingest of ‘Otkroveniye’ Audio CD[9]

The ‘Otkroveniye’ Audio CD is a collection of Orthodox Choral Music and Russian Folk Songs conducted by Irina Kozyreva and recorded in 1997-98 in St Petersburg. Although the physical carrier is a CD-R found more commonly in amateur and home recordings, the release appears semi-professional with a printed label on the CD-R and full colour liner notes. What is the most suitable method for long term curation of the audio data contained on the CD-R? An “image” file can be created containing an exact replica of the data tracks found on the entire disk in a single file. This can be written back to CD or mounted as a filesystem, allowing the file to be manipulated in the same way as the original object. However, is this the most useful format? Should we convert the files into a more commonly used digital audio format? Is it acceptable to use a ubiquitous format such as MP3 even though data is lost during compression? Or should a less common but lossless format such as FLAC be used? Although MP3 and FLAC now provide attractive mechanisms, at the time of our experiment we used the CDParanoia application to extract and convert the audio data into lossless WAV format. We took care to ensure that the sequence of the original CD was recorded and covers and liner notes suitably digitised. The ID3 tags found in MP3 files (and FLAC, among other formats) allows metadata to be easily stored with audio data. All text on the physical object was presented in both English and Russian and although we recorded bibliographic metadata in English only, recording both might be more appropriate.

Ingest of ‘Artevision’ CD-ROM[10]

The ‘Artevision: A History of Electronic Art in Spain’ CD-ROM is a compilation of Spanish electronic art produced by the Media Centre of Art and Design (MECAD) in Barcelona, Spain. This interactive multimedia CD provides the user with a unique interactive navigation system with information conveyed through videos, embedded text and scrolling images. The immersive user experience felt just as important as the information

9 The authors would like to acknowledge Irina Kozyreva & Chamber Choir, ‘Otkroveniye’ ©2000 Otkroveniye

10 The authors would like to acknowledge the Media Centre d’Art i Disseny, ‘artevision: A History of Electronic Art in Spain’, Sabadell-Barcelona ©2000 MECAD

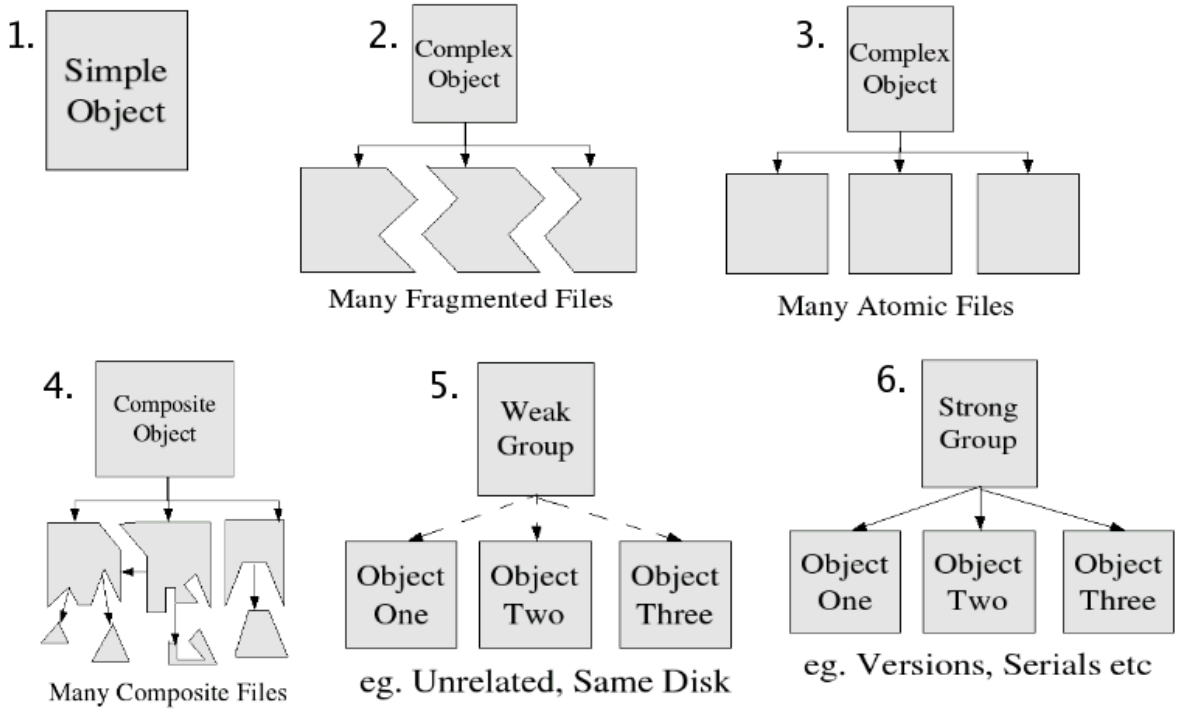
content. Viewing the content requires viewing using the Macromedia Shockwave and Apple Quicktime applications neither of which, at the time of our experiment, had native vendor support on Linux.

All textual information, both digitally and on the physical packaging, is presented in both English and Spanish. The product is well produced with the CD contained within a 210 x 146mm cardboard package with a 24-page booklet attached to the inside cover. The booklet contains an essay by the author, descriptions of the content found on the disk and production credits to complement rather than replicate the digital content. As the contents of the booklet are not available on the CD-ROM this supplementary material should be digitised and archived alongside the digital content.

The ISO 9660 standard specifies the structure of CD-ROMs. In order to create cross platform compatibility, this standard was produced according to the then most-widely used operating system, which at that time was Microsoft's Disk Operating System (DOS). Correspondingly, CD-ROM filenames were restricted to eight characters, extensions to three characters, and a maximum depth of eight subdirectories was permitted. The Macintosh operating system was not so restricted in filename usage and although developers could publish a CD-ROM volume corresponding to the native Macintosh Hierarchical File System (HFS) structure this was not compatible with the Windows platform. In response, hybrid CD-ROMs were developed to allow both HFS and ISO data to coexist on the same disk. On a Macintosh only the information contained within the HFS partition was visible and on a PC only that within the ISO partition. Shared data files adhering to the ISO 9660 standard were visible to both platforms. Since then, the Joliet extension specification has broadened the capabilities of ISO 9660, allowing longer filenames and the use of Unicode characters in filenames. Effectively, different data could be held on different partitions of the CD-ROM and without close inspection one partition could easily be overlooked during ingest. The 'Artevision' CD-ROM is a shared hybrid disk with slightly different versions available on each partition. This was not immediately apparent and goes some way to highlight the unexpected technical challenges that may be encountered.

Classification of Objects

Appropriate measures should be taken to record the context in which an object was discovered and maintain the structure and relationship between files found on physical carriers. The diagram below identifies and categorises the object structures we encountered within our sample dataset.



Simple Object
Figure 1 depicts a simple object that does not rely on additional data files to understand the information content of the object. However, it is likely that an external application is required to facilitate rendering.

Complex Object: Many Fragmented Files
Figure 2 depicts an object fragmented into many separate files which should be processed in sequence to interpret the information content. For example, a document split into one file per paragraph should be processed in sequence to be completely and correctly understood.

Complex Object: Many Atomic Files
Figure 3 depicts an object comprised of many atomic files. In many instances, it is difficult to determine whether each file should be considered as an individual object to be catalogued.

> *Example One* A CD-ROM of conference proceedings consists of a collection of reports and papers. Each report is a standalone document and does not require the information contained in another for correct comprehension. Together these files comprise a related collection as an output from a conference. Individually, each file has distinct authors, titles and dates of publication.
>
> *Example Two* An Audio CD contains a collection of tracks in a specific order. The songs are published only as a collection on the album, however each track can be listened to and interpreted individually.

Complex Object: Many Composite Files
Figure 4 depicts an object composed of interrelated files dependent upon many others stored in a hierarchical structure. Replicating this structure as a whole is required to render the object. Migrating individual files is difficult as this may break internal references.

Weak Group Construct
Figure 5 depicts objects that need a relationship between them to be recorded but without a direct relationship in content. Many distinct digital objects may be stored on a single physical carrier. Although there may not appear an intellectual relationship between these a record must be maintained for contextual and evidential value. Where physical packaging is digitised, the relationship with the information content of the original physical carrier must be maintained.

Strong Group Construct
Figure 6 depicts objects may be available in several languages or formats. It may be necessary to archive each version and if so the relationship between them should be maintained.

Practicalities of Collecting Preservation Metadata

We found manual ingest to be expensive, repetitive and likely to lead to errors. An administrator may well accidentally or intentionally declare an arbitrary, unconfirmed or incorrect value for some metadata field. We found the most time consuming aspect of ingest to be searching for and learning how to exploit utilities to extract embedded properties of binary file types. Once appropriate tools had been discovered the time required for per-file metadata creation was relatively brief. There are many freely available open source utilities which will retrieve the technical preservation metadata we require. However, these utilities are often developed for a generic or alternative goal and will only extract a certain amount of information. Tools dedicated to the automatic extraction of preservation information are needed, and it is a relief to see development work has now started on such tools in the form of Jhove[11], DROID[12] and NLNZ's Metadata Extractor Tool[13].

Many data content files (eg. image and text) specify properties such as dimensions and character sets.

---

11 http://hul.harvard.edu/jhove/
12 http://droid.sourceforge.net/
13 http://meta-extractor.sourceforge.net/

Externally recording this information is currently our most formal method of asserting the characteristics of a file and whether a file has later been rendered correctly. Further research on significant properties may be necessary. Replacement rendering tools may have limited functionality and it may be difficult to verify that an object has been rendered correctly without comparing it against the original rendering. As this comparison may not be technologically possible, descriptive information may be required to describe the object's appearance and functionality. For example, a Microsoft Word document might be written in black text with red text denoting edits and comments. A Microsoft Excel spreadsheet may contain formulae that would be lost during migration to a less fully featured format. Can we assume that future software will be able to migrate and render these correctly?

File format registries may reduce the need for repetitive technical preservation information to be extracted and stored. We found that generating the technical metadata was expensive and time-consuming. The usefulness of collecting this metadata has not yet been demonstrated and further research may be needed to assess how much metadata is needed for individual files. As long as the file formats can be identified correctly, format registries may hold authoritative information on how to access the content, either via suitable migration pathways or through emulation. This may mean per-file preservation metadata is limited to specifying the file format and version to identify the appropriate format registry entry. Identifying the version was surprisingly difficult; although generally stored in the header of the file for some formats the version information was not stored at all.

Binary files present a greater problem during ingest as their distinguishing properties cannot so easily be identified and classified. Many executable binaries are dependent upon specific operating systems and hardware configurations. These files often cannot be converted easily into more useful file formats. In these instances preservation metadata should focus on recording the configurations of hardware and software required to render and use the object. For example, consider a video game application – although migration to a new platform is possible, emulating the original environment may be more appropriate.

Understanding an object's context - what it is, why it exists - will affect its archival status. For example, one of the objects investigated was titled "Iran2002". This CD-R contains no information recording who created it, the dates of production are vague (while the timestamp of creation from the photos is present there is not information recording when the disk was made available) and there is no information stating why the disk should be archived. Understanding the context in which this information exists will determine its relevance to a collection.

Human intervention should be minimised to the creation of semantic and descriptive information, such as that specifying the creator, title, description, audience. It many situations it will be this contextual and embedded information that will be difficult for a tool to automatically extract. Heuristics should be applied to ensure enough information is recorded to allow the object to be later searched for and identified. What grammar, syntax and conventions might be most useful to support this?

Conclusions

The infrastructures for an intuitive, interconnected digital repository must be established. A complete system will consist of several components. Separate environments for content-inspection and -investigation, and metadata extraction are needed. External format and software registries are needed to provide a comprehensive catalogue of rendering tools and authoritative information on the use and functionality of each. Each component may develop and be replaced at any point, so one may best consider a digital library as an evolving service and framework rather than a static system.

Automating technical metadata extraction is essential to gather correct and accurate information, lower costs, and increase the rate of ingest. Rather than repeating metadata extraction for each file, it seems prudent to invest in file format registries providing an authoritative resource on how these formats can best be curated and the information content accessed.

Our ingest study was made more challenging by the absence of selection and appraisal policies. It was often difficult to understand why an object was valuable; why it should be preserved. While it may remain necessary to investigate the contents of each physical carrier, selection and appraisal policies are essential to

place an object in context. In addition, they may facilitate sensible disposal of irrelevant content both on the physical carriers and within the repository.

We explored a variety of physical carriers. The ingest of objects was not simply a straightforward conversion from one storage medium to another. The diversity of content makes identification, investigation and extraction of information an expensive task. The files encountered required different software applications, operating systems and platforms for investigation, and the overall ingest process was time consuming and expensive. Although we encountered some issues specific to media obsolescence, and questions of selection and identification, many of the issues we encountered are not unique to physical carriers but to digital repository processes more generally.

We found the initial process of creating an environment suitable for ingest especially expensive. Finding appropriate tools and learning how to use them when a new file format was encountered was time consuming. The cost did lower with repetition, however the overall cost of manually generating technical metadata seems prohibitive. We would promote investment to develop automated tools for this purpose. National and international collaborative efforts should facilitate this work.

Appropriate staff should only undertake those tasks for which they have been trained. For example, technical system administrators should not undertake decisions regarding selection and appraisal. In practise this will be a challenge; working with these materials needed technical familiarity and it may be necessary to cross train staff on selection and appraisal policies and procedures. Regardless of how much syntactic information can be extracted automatically, it is likely staff will need to manually generate bibliographic and contextual information.

Finally, we do not yet have a clear understanding of the practical use of technical metadata. File format and software registries may be more useful resources. Research should be undertaken on the practical use of metadata, as lowering the quantity of information collected – without affecting the quality of preservation and reuse – would significantly lower the cost of ingest.

References

[1] Seamus Ross. Digital Library Development Review, July 2003.
http://www.natlib.govt.nz/files/ross_report.pdf

[2] Derek M. Sergeant. *Synthesis of Findings from the Test Site phase of Cedars*, 2000.
http://www.leeds.ac.uk/cedars/testsitesSynthesis.html.

[3] Adrian Brown. *Managing migration: the CAMS Database and Practical Experiences in Migration*, April 2003.
http://www.pro.gov.uk/about/preservation/digital/conference/slides/brown.ppt.

[4] Reagan Moore, Chaitan Baru, Amarnath Gupta, Bertram Ludaescher, Richard Marciano, and Arcot Rajasekar. *Collection-Based Long-Term Preservation, Pages 35-46*, June 1999.
http://www.sdsc.edu/NARA/Publications/nara.pdf.

[5] The OCLC/RLG Working Group on Preservation Metadata. *A Metadata Framework to Support the Preservation of Digital Objects*, June 2002.
http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[6] Marcel van Dijk. *It Always Hurts the First Time: Experiences with transferred Electronic Records*, February 2003.
http://www.cultivate-int.org/issue9/amsterdammro/

[7] Consultative Committee for Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS), Blue Book. January 2002. public.ccsds.org/publications/archive/650x0b1.pdf

[8] National Library of New Zealand. Metadata Standards Framework- Preservation Metadata (Revised), June 2003.
http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf.

[9] Library of Congress. MARC 21 Format for Bibliographic Data (Revised). October 2001.
http://www.loc.gov/marc/bibliographic/

[10] Deborah Woodyard. *Farewell my Floppy: a Strategy for Migration of Digital Information*, April 1998.
http://www.nla.gov.au/nla/staffpaper/valadw.html

[11] Andrew Wilson. *How Digital Records are transferred to the Long-Term Digital Repository*, 2003.
http://www.naa.gov.au/recordkeeping/preservation/digital/digital_repository.html.

[12] Brian F. Lavoie. PREMIS With a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata, 2008. http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

Appendix: Sample Dataset

Title	Version	Media	Permission Rights	Creation Date	Published?	Physical Hybrid?
Clarysse Mac	-	400KB 3.5" Floppy	Granted	Mar-99	No	No
Large Scale Storage in the Web	-	CD-ROM	Granted	Apr-01	Yes	No
6th NASA Goddard on Mass Storage Systems and Technologies	-	2 x CD-ROM	Granted	Mar-98	Yes	No
The OGC Successful Delivery Toolkit	3.0	CD-ROM	Granted	Oct-02	Yes	No
Iran 2002	-	CD-R	Granted	Apr-02	No	No
A^2PAW	-	CD-R	Granted	Jan-02	No	No
Otkroveniye - Chamber Choir	-	CD-R	-	Jan-00	Yes	No
artevision - A History of Electonic Art in Spain	-	CD-ROM	-	Jan-00	Yes	Yes
Le Mundaneum	-	CD-ROM	-	May-98	Yes	No
The Mission Transcript Collection	-	2 x CD-ROM	Granted	Jul-02	Yes	No
Digital Culture: Maximising the Nations Investment Word 6.0 Files	-	1.44MB 3.5" Floppy	Granted	Feb-99	No	No
Managing Electronic Records in an Electronic Work Environment	-	1.44MB 3.5" Floppy	-	May-96	Yes	No
Core Resources for Historians	1.0	CD-ROM	-	May-98	Yes	No
Cornaline de l'Inde	-	CD-ROM	-	Jun-00	Yes	Yes
Archives of Poland	-	CD-ROM	-	Jun-00	Yes	No
Year of the Lithuanian Book	-	CD-ROM	-	Jan-98	Yes	No
Towards a Knowledge Based Society	-	CD-ROM	Granted	Oct-01	Yes	Yes
Needs Assessment ICT Questionaire	-	1.44MB 3.5" Floppy	Granted	Aug-00	No	No
Mailing Labels IT Study HLF	-	1.44MB 3.5" Floppy	Granted	Sep-97	No	No
PINS2	-	1.44MB 3.5" Floppy	Granted	Sep-98	No	No
ACLS Hum, Word 6.0.1 Macintosh; Acls.txt, ASCII	-	1.44MB 3.5" Floppy	Granted	Unknown	No	No
Handlist.ulm, WP5.1 format for DOS/Windows 3.1	-	1.44MB 3.5" Floppy	Granted	Jan-89	No	No
Preston's Illustrations of Masonry	-	CD-ROM	Granted	May-01	Yes	No
Cantar de Mio Cid	-	CD-ROM	-	Jun-98	Yes	Yes
Libro de los gorriones	-	CD-ROM	-	Jun-99	Yes	Yes
ecalsi 2002 Workshop: 22-23 Octubre 2002	-	CD-R	-	Oct-02	No	No
The First Cincinnati Haggadah: An Interactive Facsimile Edition	-	CD-ROM	-	Jun-00	Yes	Yes
Das Alteste Burgeraufnahmebuch der Reichsstadt Regensburg	-	CD-ROM	-	Jan-97	Yes	No
Tinguj E Zera Te Shekullit: 1900-2000	-	CD-R	-	Jan-00	Yes	No
Juvenate: an interactive narrative	-	CD-ROM	-	Feb-01	Yes	No

With Open Eyes	-	CD-ROM	-	Jan-95	Yes	No
Corpus of Romanesque Sculpture, Sussex	-	1.2MB 5.25" Floppy	Granted	May-91	No	No
Letters & Abstracts-1	-	1.2MB 5.25" Floppy	Granted	Unknown	No	No
Activate Disk	-	1.2MB 5.25" Floppy	Granted	Unknown	No	No
(Unlabelled Collection)	-	1.2MB 5.25" Floppy	Granted	Apr-87	No	No