Exploring Strategies for Digital Preservation for DSpace@Cambridge

Grace Carpenter
Massachusetts Institute of Technology
Cambridge, Massachusetts

Jim Downing
Cambridge University
Cambridge, England

Abstract

Preserving the functionality of digital objects (as opposed to just the bitstream), whether text, images, or scientific data, is a dauntingly complex task. The DSpace@Cambridge project, a joint project of Cambridge University Library and MIT Libraries, in partnership with Cambridge University Computing Service, has among its goals to further develop DSpace in the area of digital preservation. With this project we hope to demonstrate the possibilities of implementing digital preservation within the context of existing institutional repositories, and also to cover new ground in some small areas of digital preservation. By evaluating existing tools for integration into DSpace, and perhaps creating tools that can be widely re-used, we can provide a concrete example of their use. The digital preservation project is a work in progress, and in this paper we report on our current status and also on our plans for the remaining portions of the project.

There are three main themes in the DSpace@Cambridge research: process automation, preservation planning, and a specific file format investigation on ArcView export file format.

Automation

Digital preservation can only be carried out if there is a solid foundation on which to build it. The accurate identification of a file's format, the validation of the file against type specification, and extraction of technical metadata are all part of this foundation. The National Library of New Zealand (NLNZ) and Harvard University Library have both recognized these issues and recently created tools that perform format identification on digital objects, as well as automatic extraction of technical metadata from them, and in the case of Harvard's JHOVE[1], format validation. We have examined inconsistencies that have arisen in the ingest process at MIT's DSpace repository and then evaluated the available tools for solving the problems we've encountered.

File Identification: The Issues

There are several ways to determine a file's format: by file extension (MIME type); by reading the "magic numbers" at the beginning of the file; by parsing an object in its entirety and determining if it's a valid instance of certain format; or by using some combination of the above. DSpace currently identifies objects by the file extension, but offers the possibility of a manual override. A comparison of totals in MIT's DSpace repository (as of January 2005) of the number of files by recorded file format vs. totals of the number of files by file extension showed some significant discrepancies between the two counts, suggesting that a submitter's knowledge of a file format does not always match the MIME type. Below is a sample of some of the more notable discrepancies:

Recorded file type:	Total files:	Associated file extensions:	Total files:
pdf	5879	.pdf, .pdf-3	5867
xml	1	.xml, .dtd	0
jpeg	34	.jpg	29
gif	56	.gif	59
postscript	1462	ps	1543

The query of file extensions also showed numerous obscure file extensions, some of which were officially identified as "unknown", although they presumably would have fit into the better defined categories of "text" or, minimally, "bytestream". The confusion on file formats that exists within MIT's own repository illustrates the great need

for even the most basic of preservation support services, such as better identification on ingest.

File Identification: the Options

We considered three tools for identifying file formats: the Unix "file" command; the NLNZ Metadata Extractor Tool (reviewed in a beta version)[2], and Harvard University Library's JHOVE (reviewed in version 1.0 beta 3). The Unix/Linux "file" command identifies non-system files based on the "magic number", an "invariant identifier at a small fixed offset into the file"[3] in binary files, or on some of the contents of text files. Although this would be a very simple solution, there are several drawbacks to it, including having to run the command at the command line (as opposed to calling an API), platform dependency, and a simplistic approach that doesn't take into account the possibility of invalid files. Both JHOVE and the NLNZ tool perform not only file identification but also metadata extraction, which is an essential part of this project. However, NLNZ uses file extensions to identify an object's format, and therefore doesn't provide DSpace with the necessary level of certainty beyond what we already have for identification. In contrast, JHOVE's default method of identifying a file's format is to do a complete parse of the object to determine if it is a valid instance of a particular format, rather than simply a purported instance of it (i.e., has the specified file extension or magic number). JHOVE does also allow the user, though, to specify identification by magic number only.

How to definitively identify an object's format within DSpace is still an open question; we hope to make DSpace configurable so that each DSpace instance can determine file formats with a method deemed appropriate by that institution. The general consensus, though, is that using file extensions alone is incomplete and too susceptible to error. Therefore we've chosen to focus our efforts on integrating JHOVE into DSpace, as it appears to provide the most reliable method of file identification. In addition, JHOVE provides technical metadata extraction, which is discussed below.

Refinement of File Type Definitions

A crucial issue in identifying file formats is the creation of a file type definition that is a good deal more precise than MIME type; for instance, one that includes, at a minimum, a version number where relevant. A number of initiatives in recent years have come a long ways in addressing this issue, such as the Global Digital Format Registry (GDFR)[4] as prototyped by the FRED[5] project, and the UK National Archive's PRONOM[6]. DSpace@Cambridge hopes to create a base implementation of an institutional repository using format definitions from FRED, and to contribute our experiences to the expansion of the GDFR.

Extracting and Using Technical Metadata

Our experience has shown us that users (content providers in this case) are often somewhat uncomfortable with the concept of technical metadata, and only occasionally share the digital curator's motivation to enable long term preservation. Therefore it is essential to lower the barriers of adoption to digital preservation by extracting metadata automatically and thereby reducing the level of human effort required in the ingest process. In addition, the enormous rate of some content provision requires an entirely automated process from creation to ingest. This includes ensuring that relevant technical metadata is present before ingest, and can be automatically extracted. We hope to begin our testing of JHOVE integration into DSpace in early spring of 2005.

Once past the initial hurdle of extracting technical metadata, there is still the question of how to store it. Currently DSpace does not require technical metadata for ingested objects, and cannot provide search or browse facilities for it. Therefore a part of this project focuses on how to best integrate this metadata into an existing system. Our plan is to transform the extracted technical metadata into the METS format and to serialize it into an XML file along with the data file.

There are many uses for technical metadata; we expect to focus our efforts initially on using it to create preservation strategies. Members of the JHOVE team have pointed out that each digital object must be examined on ingest within the context of what the user finds to be the essential aspects of that particular object, in addition to the wider context of that class of objects. The creators of JHOVE have described what their tool does as 'characterization':

"Format *characterization* is the process of determining the format-specific significant properties of an object of a given format, e.g.: 'I have an object of format *F*; what are its salient properties?'"[7]

Although JHOVE does not currently parse Microsoft Excel files, we will use Excel here for the sake of example in discussing how the characterization of a file could be useful in creating different preservation strategies for objects of the same format. One of the difficulties of preserving Excel spreadsheets is that Excel is used across many disciplines for many purposes; it would be almost impossible for users to agree on what is the preservable "essence" of a spreadsheet. For some users it may function solely as a data repository, in which case its essence can be probably be preserved as a .csv file. For others it may function as a presentation tool, and the information in it could be transmitted equally well as a .pdf or .html file. For yet others the analytical functions contained in the cells may be as crucial as the visible information, and only the spreadsheet itself can do justice to the multiple layers of information. In addition, some spreadsheets contain macros that transform them into applications, rather than static containers of data and functions. The ability to characterize a particular spreadsheet format (such as Excel) would enable preservation specialists to determine a preservation strategy that is appropriate not just to that format, but to its intended audience for its primary purpose

Ensuring Content Fixity

Providing the repository administrator with an assurance of content fixity is another essential part of building a preservation foundation in a digital repository. DSpace currently enables content fixity checking by storing an MD5 digest of a file as a checksum in the database. In collaboration with University of Rochester we are adding automated ongoing verification of these checksums, and tools that will assist the administrator in resolving mismatches.

Calculating digests in this way is I/O bound, and an extremely modest processing capacity could easily saturate a typical asset store's I/O. Consequently we need to avoid performing these checks when the repository is in use, and will limit execution to a fixed period of time (rather than fixed number of files) each day. This leads to a trade off between allocating time to the checksum checking process (rather than other periodic tasks such as database maintenance and backup) and the time taken to perform a complete iteration of the repository (which could guide the repository's backup strategy).

For example, the http://www.dspace.cam.ac.uk/ uses an asset store based on an EonStor RAID with a practical sustained maximum bandwidth of 90MB/s. At capacity the store will contain 10TB and will require over 32 hours of dedicated access to calculate checksums for the entire store.

If it is assumed that the only access to files is performed through DSpace, then a checksum check could only fail due to file corruption. However, in practice, administrators occasionally bypass the DSpace API and handle files manually, and therefore it was decided to consider a variety of possible causes of a checksum mismatch. DSpace administrators will be provided with an extensible problem tracking and resolution tool capable of handling conditions not anticipated by the original development team.

Preservation Planning

One of the more frustrating aspects of creating preservation strategies for different formats is the lack of a centralized information base on what other similar institutions are doing. Although published papers often report on groundbreaking preservation initiatives, it can be difficult to get a clear picture of preservation options—such as migration tools—that are available to those without the resources to create their own in-house systems.

Building on work on format action plans done at the Florida Center for Library Automation as part of the DAITSS project[8], we hope to create templates for strategies that will both help institutions get started in their planning, as well as provide a means for sharing information and strategies among institutions. Our ultimate goal is to create a system of machine readable preservation strategies that can evolve to support future rendering processes, and yet retain enough information that such processes can be human validated. Although the initial aim will be on migration, it is hoped that the technique can be extended to emulation and Universal Virtual Computer approaches. Our hope is to prove this preservation strategy approach by writing a migration tool for one or two formats capable of supporting migration on ingest or migration on-the-fly.

Specific Format Investigation

The majority of existing preservation efforts have focused on text and raster image formats. We hope to extend the focus of these efforts by developing an in-depth knowledge of vector GIS data and developing strategies specifically for the ArcInfo and ArcView file formats. Extending JHOVE for a GIS format is a possibility, as is developing a migration tool for GIS data. It is hoped that some of the learning points from this investigation will be relevant to other vector formats (CAD/CAM, vector graphics).

Conclusion

In addition to facilitating the preservation of digital objects in DSpace, this project hopes to serve as a testing ground for some of the issues involved in digital preservation. Due to its large institutional user base and its Open Source community, DSpace provides an unusual opportunity to try out different tools and strategies and to garner feedback on their utility from users and developers. We see this also as an opportunity to test out some of the ideas and the data model put forth in the Global Digital Format Registry, and in turn to make a contribution to the ongoing conversation on digital preservation.

References

1. http://hul.harvard.edu/jhove
2. http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction
3. File command man page (http://www.google.com/search?hl=en&q=unix+file+man)
4. http://hul.harvard.edu/gdfr/
5. http://tom.library.upenn.edu/cgi-bin/fred?cmd=ShowDocu&&id=about
6. http://www.nationalarchives.gov.uk/pronom/
7. http://hul.harvard.edu/jhove/
8. http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf

Biography

Grace Carpenter is a programmer in the Digital Library Research Group at MIT Libraries and a contributor to the DSpace@Cambridge project. Her previous experience includes software development and systems work in both the financial world and non-profit organizations. She holds a BA from Barnard College of Columbia University.

Jim Downing is a programmer on the DSpace@Cambridge project, and a committer on the

DSpace software project. His previous experience includes software development and consultancy in domains from knowledge management to financial services through science and engineering research. He holds an MEng(hons) from Cambridge University.

Exploring Strategies for Digital Preservation for DSpace@Cambridge

Grace Carpenter
Massachusetts Institute of Technology
Cambridge, Massachusetts

Jim Downing
Cambridge University
Cambridge, England

Abstract

Preserving the functionality of digital objects (as opposed to just the bitstream), whether text, images, or scientific data, is a dauntingly complex task. The DSpace@Cambridge project, a joint project of Cambridge University Library and MIT Libraries, in partnership with Cambridge University Computing Service, has among its goals to further develop DSpace in the area of digital preservation. With this project we hope to demonstrate the possibilities of implementing digital preservation within the context of existing institutional repositories, and also to cover new ground in some small areas of digital preservation. By evaluating existing tools for integration into DSpace, and perhaps creating tools that can be widely re-used, we can provide a concrete example of their use. The digital preservation project is a work in progress, and in this paper we report on our current status and also on our plans for the remaining portions of the project.

There are three main themes in the DSpace@Cambridge research: process automation, preservation planning, and a specific file format investigation on ArcView export file format.

Keywords

Digital preservation, technical metadata, preservation strategy