DSpace@Cambridge: implementing long-term digital preservation

Tom De Mulder

August 30, 2005

Abstract

DSpace@Cambridge is an institutional archive set up to deal with the long- term preservation of information in a wide range of formats over an indefinite period of time.

In this paper we look at some long-term digital preservation strategies, as they are currently implemented in our archive.

We describe the value of documentation of file format specifications for future data accessibility. We examine the impact and usefulness of constant concurrent data migration to several different formats.

We illustrate our approach with case studies of applying these principles to chemical and archeological scientific data.

1 Introduction

A common approach to digital preservation is to try and keep an archive's contents accessible, through format migration, emulation and so on. A lot of research effort is going into this, and is certainly worthwile.

However, what if one was to think really long-term? Imagine a moment in the far future, where someone unearths a digital archive that hasn't been actively maintained for decades. How can we help this future "digital archaeologist" to reconstruct the data in the archive?

One approach is to provide documentation, explaining the workings of the archive as well as the structure of its contents, and the structure of the file formats used.

Another aproach is to provide the same content in different formats, so increasing the chance that at least one copy of the data will be readable. If necessary, this can then be used as a template to help reconstruct the original.

In our methodology, we assume this future digital archeologist to be moderately intelligent. We assume that any conclusions we can draw from patterns spotted in data, (s)he can come to, too.

We also assume that the contents of the archive are available as a stream of zeroes and ones, so we won't cover hardware obsolesence but take it for granted that the future archeologist is presented with a set of accessible bits. With some basic pattern matching, it should become apparent how this data is structured in files, and that some of these files contain (english, in our instance) text.

Much of the content of this paper originated in some of the author's work trying to reconstruct data from old/obsolete systems. The approaches suggested here would have made that task a lot easier or, sometimes, at all possible.

2 Documentation

The importance of documentation for digital archives is well understood [1] [2] when it comes to documenting the data creation process. Often, it is defined as follows: "The information provided by a creator and the repository which provides enough information to establish provenance, history and context and to enable its use by others." However, we suggest going one step further and also documenting the archive itself, and the standards in use while the archive was active.

These can all help reconstruct the archive and its contents. While the documentation describing these standards is ubiquitous right now, there is no guarantee that it still will be in the future. For any scenario where someone in the far future has to reconstruct the archive, we can't assume that any or all of these documents will still be around. The only safe place for them to be, if we want them to be found, is inside the repository alongside the data they describe.

It is of course important that the documentation itself is easily readable, lest the key to the treasure be locked in with the treasure itself. The best approach in this case is to store documentation as plain text files, because english ASCII text is easily recognised as such. In fact, when a future digital archaeologist goes searching through the archive, the documentation data would probably be found first, because ASCII text's binary representation is distinctive and easily deciphered.

This is an important consideration for someone who might be dealing with a random unstructured set of zeroes and ones, rather than a fully operating repository including its supporting software.

We also include supporting documentation for file formats. These explain how files are structured, so their contents can be recovered. This is easily done for open standards, where documentation is widely available and licenses allow for these standards to be disseminated. It is, however, very hard or even impossible for closed formats, e.g. MS Office documents. This is a very good reason to use open formats where possible.

As mentioned earlier, we also want to include

[1] AHDS History Documentation Guidelines: http://hds.essex.ac.uk/docguide.asp

[2] Cornell Digital Preservation Tutorial: http://www.library.cornell.edu/iris/tutorial/dpm/

documentation on technology conventions in use during the lifetime of the archive. One obvious source for this are the IETF's Requests For Comments (RFCs) [3]. These describe most of the way the internet currently works, which is useful background information for the way current archives work. And, while they certainly can be found mirrored in hundreds of places around the internet at the moment, including them doesn't take up much archive space and means we can safely assume that they will be available when needed.

Finally, there's the documentation on the archive itself. Where possible we also include the source code to the software running the archive. Because, while it's unlikely that the programming languages used now will still be in use in the far future, they tend to be logically structured and may relatively easily be reconstructed when researched.

3 Format concurrency

In full: "Multiple format concurrency", by which we mean: storing the same data in multiple different formats, one alongside the other. For example, digital images could be converted into TIFF, JPEG, PNG, even PDF, and all these versions stored in the archive. Sound files could be in PCM ("wave") format, mp3, ogg and AAC.

The advantage of digital format migration, after all, is that any action on a copy of the original doesn't alter the original. So, even if a migrated copy of a data file has lost some information (through compression or bit loss), this does not affect the original.

When these formats are mostly functionally equivalent, then they can be the digital equivalent of a Rosetta Stone, and help with reconstructing all the information from the original file.

As an example, should a spreadsheet be converted to a set of high resolution images, one for each slide, then some functionality is lost because the arithmetic and values behind the spreadsheet are no longer present. However, this "snapshot" of the data can be a tremendous aid for someone trying to reconstruct the original file: (s)he will know what the result should look like.

Hence the comparison with the Rosetta Stone, a granite stone slab containing the same Ptolemaic decree in greek, Egyptian demotic script and Egyptian hieroglyphics. Although the greek translation of the text expresses concepts fundamentally differently from the Egyptian hieroglyphics, it provided enough clues for Jean-Franois Champollion to decipher the latter. The importance of the greek text resided in its containing broadly the same information, and served as a rough template to reconstruct the actual content.

Figure 1: The Rosetta Stone

In a digital context, this conversion is easy to do at ingest, or at a later stage should new file formats become available that can carry the information. Ideally every file should have at least one copy in the archive in a format for which the archive itself contains format documentation, again improving the chance of future digital archaeologists recovering the data.

4 Conclusion

By and large, digital storage space gets cheaper over time. Adding documentation to an archive is a cost-effective way of safeguarding its contents for the future. Multiple format migration has a larger

[3] http://www.rfc-editor.org/

impact on the archive's technical requirements, but could prove a valuable aid in reconstructing digital assets in obsolete formats.

5 Case studies

5.1 Horse Paleopathology data

One of the collections in DSpace@Cambridge contains horse paleopathology data. [4] These are research findings from excavations in China in 2004, and consist of images of horse bones and associated measurements.

The original photographs were made on traditional celluloid film and then scanned to convert them to a digital format. The original scan is stored in a dark archive, while a version that has been photographically cleaned up to improve visibility and contrast is stored in the public archive. This image is in Adobe's proprietary Photoshop (PSD) format, but was, before import, converted to JPEG. Two resolutions were chosen for the conversion: one full resolution version to aid with format reconstruction, and one low resolution version to speed up current access to the archive contents. It is planned to also include a TIFF format of each image file in the future.

Data is virtually meaningless without metadata, however. Our archive by default only supports Dublin Core metadata, and where possible these fields were populated. Because Dublin Core wasn't made to deal with data describing horse bones, we decided to store this data alongside the image files in the archive as a set of plain text files.

5.2 Archeological data

The Kilise Tepe project [5] involves a series of excavations in the Gksu Valley in Turkey. The Project's archive in DSpace consists of all the scanned photographs (as low resolution jpgs), the 12 interconnected FileMaker databases (in their original format and as csv files), pdf versions of the illustrations and other data in the form of tables, graphs, etc. from the monograph Excavations at Kilise Tepe, 1994-1998. Also archived are the original scanned photographs, plates of drawings and plans, and the text of the monograph, although these are not available on open access for the foreseeable future, so as not to undermine the commercial viability of the publication.

Where possible, the documents were converted into similar formats. Word documents were converted to RTF and PDF, and images to PNG.

Spreadsheets were converted from the original Excel format to tsv (Tab Separated Value plain text files), losing some funcationality, such as markup and formulas, but maintaining the actual data content.

5.3 Chemical information

DSpace@Cambridge serves as a publishing tool for hundreds of thousands of molecules from the Unilever Centre for Molecular Informatics' "World Wide Molecular Matrix" project[6] [7]. In this case, the data itself is already in Chemical Markup Language, a variant of XML. The standard itself is open and, because it is encoded in plain text, easily parsed.

While in this case we don't migrate the data to other formats to be stored alongside these originals, we do provide documentation not only on the CML file format, but also on the software used to import these files into our archive.[8]

[4] http://www.dspace.cam.ac.uk/handle/1810/31293
[5] http://www.dspace.cam.ac.uk/handle/1810/31289
[6] http://www.dspace.cam.ac.uk/handle/1810/724
[7] http://wwmm.ch.cam.ac.uk/
[8] http://www.dspace.cam.ac.uk/handle/1810/52544