Practical Preservation

The MLA Yorkshire Archive at West Yorkshire Archive Service

This case study outlines the first steps in practical digital preservation taken at West Yorkshire Archive Service. The closure of the regional sector support agency, MLA Yorkshire, was seized upon as an opportunity to derive simple, 'good enough' solutions to the diverse preservation issues posed by a real archive containing digital materials. Research into available software tools and preservation services led to the development of new procedures and documentation for processing born-digital content. Contrasting this workflow with the steps taken in regard to paper materials in the same organization's archive, this project also demonstrates the enduring relevance of traditional archival skills to getting started in digital preservation.

Introduction

West Yorkshire Archive Service (WYAS) is the joint local government archive service for the five metropolitan coucils in West Yorkshire (Bradford, Calderdale, Kirklees, Leeds and Wakefield). WYAS collects and makes available to the public local historical records of all kinds dating from the twelfth century to the present day.

MLA Yorkshire was one of nine regional strategic development agencies for the museums, libraries and archives sector which were wound up at the end of 2008, due to a restructuring of the Council on Museums Libraries and Archives (MLA). MLA Yorkshire was quite independent of the MLA. A charitable organization, MLA Yorkshire and its precursor, the Yorkshire Museums, Libraries and Archives Council (YMLAC), had a longer ancestry, developing out of the former Yorkshire Museums Council (YMC) and its predecessors.

WYAS agreed to accept the archives of MLA Yorkshire, which included management board minutes, annual reports and accounts on paper dating back to the foundation of the Yorkshire Museums Council in 1963. MLA Yorkshire did not initially envisage handing over any digital material to the West Yorkshire Archive Service. For WYAS, however, the contents of MLA Yorkshire's server represented a largely risk-free testbed for an embryonic digital preservation programme, since the principal governance documents of the depositing organization were duplicated in hard copy. A functional analysis of MLA Yorkshire's operations also revealed that the public face of the organization was primarily a digital affair, via a

E-bulletin from MLA Yorkshire: Left as it appeared in its native web-based environment Right once exported into PDF-A

website and a series of e-mail bulletins. These records were particularly at risk, since MLA Yorkshire was paying for their hosting service online and they did not exist in any equivalent analogue form. Since the Archive Service's preparations for digital archives had been mostly predicated upon receiving office-type documents, these web-based materials represented a different kind of challenge.

The aim of the project, therefore, was simply to endeavour to keep MLA Yorkshire's digital output accessible locally, at least in the short-term. WYAS hoped to build on the experience with the MLA Yorkshire archive, but did not seek to forecast what this future might look like. The knowledge gained from working with the MLA Yorkshire archive would be valuable whether WYAS sought eventually to develop as a digital repository or wished to negotiate terms with a third party digital preservation supplier.

> Nobody has the perfect answer to digital preservation for every case. If we try we may fail; if we don't try we will certainly fail.

Staffing and Skills

The practical work on the project was undertaken by a single member of WYAS staff, within a framework planned by a small internal working group made up of four archivists, with IT support from West Yorkshire Joint Services. WYAS also gratefully acknowledges technical suggestions put forward by MLA Yorkshire's IT support company, and from experts in the digital preservation community. An early lesson learnt was the need, when negotiating a digital deposit, to cultivate a good working

relationship not only with the depositor, but also with the depositor's IT experts.

> An early lesson learnt was the need to cultivate a good working relationship not only with the depositor, but also with the depositor's IT experts…Good interpersonal skills are important in all preservation endeavours.

Surveying the archive

Negotiating the deposit took place over a very short timescale of just six weeks. This proved to be a real challenge in regard to the digital content in particular, but one with which the staff of any collecting archive, such as WYAS, will already be familiar. With the exception of the website and e-mail bulletins already mentioned, most of the digital material was held on a single server and the records themselves were found to be predominantly in current Microsoft Office formats.

The staff of MLA Yorkshire was asked to help prepare their filing (both electronic and paper) for the transfer. This included copying photographs held on CDs onto the server, boxing up paper records and deleting personal e-mails. Confidentiality was a significant concern requiring sensitive management, so that any passwords or authentication routines which might prevent the processing of digital content could be detected and removed before the organization closed. MLA Yorkshire's IT support personnel also helped to prepare directory and item listings of the contents of the server, which allowed the archivist at WYAS to begin appraisal and initial analysis of the archive's structure before the transfer. In total, WYAS prepared to receive 1.4 cubic metres of paper documents dating from 1963 to 2008, and around 80GB of digital material, c2002 to 2008. It was anticipated that both paper and digital sets would be heavily weeded during appraisal post-transfer.

Collecting digital material

With the MLA Yorkshire server to be decommissioned and sold under Charity Commission rules for the winding up of organizations, the most urgent consideration became that of finding some means to copy the digital material in order to transfer it to the Archive Service. The only digital deposits to have been received previously by WYAS had been small and had arrived on personal, portable media, such as floppy disks or CDs. Never before had WYAS staff had to go out to collect digital content. A USB powered external hard drive, with a terabyte of capacity, was purchased for this purpose. A small dowry was negotiated with the archive to cover this cost, which also paid for acid-free boxes for the paper documents.

Some large database applications, felt to be of limited long-term historical value, were deliberately excluded from the data collection exercise at this stage. Since these included the organisation's routine financial and client management control systems, this fortuitously and coincidentally dealt with MLA Yorkshire's outstanding data protection and confidentiality concerns. Nevertheless, the quantity of digital content to be copied was large. Fortunately, most IT support staff will be familiar with a variety of file synchronization tools, used for network maintenance, which are equally applicable to manipulating and copying digital files for archival purposes.

Classical archival theory puts much emphasis upon maintaining 'original order' as one means of establishing the records' authenticity. One advantage of the digital world over the analogue is the ease with which an entire filing structure can be duplicated, wrapped, and authenticated by means of computer-generated 'checksums'. A free 'Lite' version of a digital forensics software tool was used to perform the data capture onto the portable hard drive, generating unique checksum values (using both MD5 and SHA-1 algorithms) for each file and each package.

Documentation

Although WYAS had developed detailed new documentation for the receipt of digital archives, this had not been devised with such a large deposit in mind. The resolution was to use the formal documentation to record summary details for the entire digital deposit (for instance, the technical details of the server from which the records had been copied; the current version of Microsoft Office in use at MLA Yorkshire at the end of 2008), whilst relying upon software tools to extract technical details for each file automatically (such as DROID for file format identification). DROID is a simple and free to use program provided by the National Archives which compares the contents of a disk drive to a large online database of file formats (called PRONOM)*. This helps you to identify the formats of files stored on disk and advise about any immediate errors or possible problems. Knowing the numbers of each file format present and how they are likely to behave through time is a great help in deciding how best to manage a collection. The information derived from the forensics software application during the data capture was also imported into a spreadsheet for simple analysis, such as identifying

duplicate items, and to give a rough count of the number of items of particular format types (for example, Word files).

Processing

Once returned to WYAS, the task was essentially to reverse the data capture process – copying the files off the external hard drive, and recreating the filing structure from the images created by the digital forensics software. Since the external hard drive had potentially been compromised by connection to the depositor's system, the device was plugged into a standalone computer isolated from the WYAS network to ensure that no virus could be transferred inadvertently. All initial processing took place on this standalone 'ingest' machine.

> Classical archival theory puts much emphasis upon maintaining 'original order' as a means of establishing the records' authenticity. One advantage of the digital world is the ease with which an entire filing structure can be duplicated, wrapped, and authenticated.

The incoming material was first virus checked. A free anti-virus package was chosen: partly on cost grounds, partly because it was not the application used for virus-checking on the WYAS network, the logic being that the digital archive would be checked a second time for virus infection, using another anti-virus package, when finally transferred to longer-term network storage. The initial virus check was run twice. Once, immediately the data was copied from the external hard drive, and again about a month later with updated virus definition files.

Appraisal, description and metadata

The spreadsheets created during the transfer process also served as temporary finding aids to the digital part of the archive, as they listed digital file titles, file paths and creation dates alongside more technical details such as file size, file extension and checksum value. In this sense, they were the digital equivalent to the box lists which had been manually put together to accompany the paper documents in the collection.

These spreadsheets raised quite a number of appraisal and cataloguing issues. It was evident that the collection as transferred included a considerable quantity of reference copies of policy documents which had evidently been downloaded from the Internet, plus the usual software configuration files and example data, such as templates and clipart. None of these digital items could strictly be described as part of the MLA Yorkshire archive, and if WYAS decided to follow a migration strategy of digital preservation, this type of material would be stripped out. However, these files might be required if an emulation policy was chosen. There was also considerable duplication: should this be maintained in order to retain context and original order, or was it more confusing to retain identical documents in different parts of the collection? If the decision was made to weed out duplication, which copies should be deleted and which retained?

There was also the question of those document series which were mirrored in both digital and paper formats: could the digital copy be safely disposed of where a paper version existed, particularly if the paper version was signed (as with annual accounts, for example)? On the other hand, would it be preferable to retain a digital surrogate for future access purposes?

How should the digital material best be represented in the Archive Service's catalogue database, constructed according to the archival descriptive standard ISAD(G)? Indeed, was a separate finding aid even required for the digital part of the collection, given the level of detail already available in the spreadsheets?

Web-based content

The MLA Yorkshire website was due to be replaced on the day the organization closed with a single holding page re-directing readers to the national MLA website.

Snapshot of the MLA Website in December 2008

Several web harvesting services are available, and WYAS tested one of these to take a snapshot of the website in the last week of MLA Yorkshire's existence. The initial tests produced only a series of errors, and so the trial became a useful introduction to some of the difficulties of web archiving. The importance of good interpersonal skills to preservation projects again became evident when

special access for the web crawler had to be negotiated through the London-based MLA web manager.

Described to WYAS staff as 'the main organ of getting information out' of MLA Yorkshire, the e-mail bulletins proved problematic from a preservation point of view. The only complete set of bulletins issued was held by a commercial online marketing service based in the United States. This service had enabled MLA Yorkshire staff to draft newsletters using customized templates, and then send them out of selected mailing lists held within the system. The service then saved a considerable amount of metadata about when the bulletin was sent, how many people it was sent to, how many e-mails bounced, and the 'click-through' rate for hyperlinks embedded in the e-mails.

An 'archive' facility was available within the online system, but this was an extra, chargeable service. In any case what WYAS really wanted to achieve was an offline copy of the content. The option of forwarding all the e-mails again to a WYAS e-mail account was rejected, since although this would keep the content of the bulletins available, it would destroy the associated contextual metadata, such as the date the e-mail was originally sent and the details about the readership. Eventually, a pragmatic solution was found which involved exporting the 'printable' view of each bulletin into PDF-A format, with the accompanying metadata saved in the same way. This did involve some loss of formatting detail, such as colours and the functionality of embedded hyperlinks (many of which were by now broken links in any case), but was felt to be a reasonable compromise to enable the information content and metadata to remain available.

Summary

The MLA Yorkshire project is an example of a small organization's initial attempts to move beyond abstract theory and policy statements and take some practical steps forward with regard to digital preservation.

This archive would not have fallen into the collecting remit of any of the established digital repositories, not least because of its mixed digital and paper content. For WYAS, however, processing a hybrid archive like MLA Yorkshire was particularly instructive, as it brought to light both similarities and differences in the workflow required, and WYAS was able to adapt from current procedures as well as create new documentation specific to the digital domain. WYAS gained practical knowledge of available software tools and preservation services, and realized that there was little point in developing elaborate procedures to cope with obsolete digital formats, when the majority of material was received in current formats. And as a result, WYAS have been able to respond positively to several enquiries about the digital materials in the MLA Yorkshire archive.

There are, of course, many decisions still to be made. This case study has presented steps taken to ensure that the MLA Yorkshire digital archive remains accessible, locally, in the present. It is less certain whether WYAS has the technical and infrastructure capability - or desire - to curate the MLA Yorkshire digital content indefinitely. But that was not the point of the project. The MLA Yorkshire archive remains accessible and the transfer and preservation actions which have been taken upon it have been documented. All preservation endeavours in any case rely upon partnerships: partnerships between archivists and technical experts, and partnerships across the generations of record keepers. WYAS themselves hope to build on this project, but also to inspire other small archive services to take their first steps in practising practical preservation, helping them to fulfill their part in the preservation chain.

*For more information about DROID and PRONOM, see: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

This case study was prepared by **Alexandra Eveleigh** of **University College London**, and formerly of **West Yorkshire Archive Service**, with the assistance of **MLA** and the **Digital Preservation Coalition**. It was made possible with funding from **JISC**. October 2010.