**The Application of Preservation Planning Templates to a Personal Digital Collection – Research Notes**

Paul Wilson, 2 Soames Close, Lavendon, Olney, Bucks, MK46 4EJ, UK E-mail: [pwilsonofc@btinternet.com](mailto:pwilsonofc@btinternet.com)

**Abstract**

**This research note summarises the experiences of the author when applying Preservation Planning processes and templates defined in a previous stage of work.**

**Keywords**

Digital Preservation Planning; Preservation Planning Process; Document Templates; Personal Collections

**Introduction**

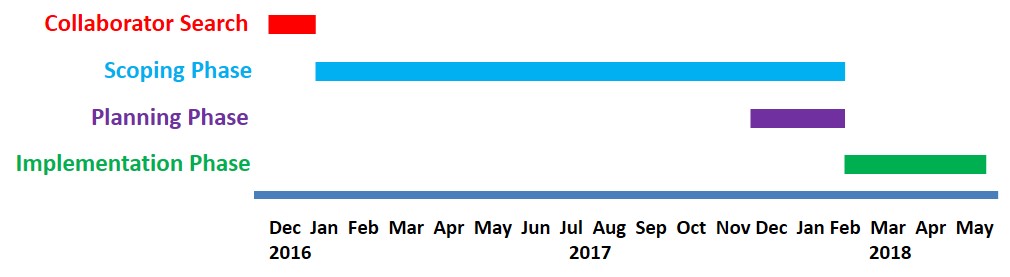
My lifetime Collection of work documents (PAWDOC) was initially established in 1981 to gain practical experience of the emerging field of Office Automation when I was working in the newly constituted Office Systems Division at The National Computing Centre in Manchester [Wilson, 2001]. Once established, the Collection became, de facto, an integral and essential part of my working life. Most of my work documents - both paper-based and electronic - were stored in it up to my retirement in 2012 - and a few more since then. All the paper documents have been scanned to electronic files. The index to the Collection contains over 17,000 entries, and each entry represents one or more of the 29,000 electronic files.However, in recent years I became increasingly concerned that files in particular formats in the Collection may not open; and that the cost of purchasing and implementing upgrades to the Index software and Document Management application was becoming prohibitive. So, in 2014, I set out to find a simple digital preservation workflow that I could apply to a personal collection like mine. The results of that research are documented in the DPC Research note of April 2016 [Wilson, 2016] which also included templates for the following documents:

* Scoping document
* Preservation Project Plan Description
* Preservation Project Plan Chart
* Preservation Maintenance Plan

In 2017, I set about applying the process and templates to the PAWDOC Collection. This paper describes what was experienced and identifies improvements that have been made to the templates.

**Overview and Timeline**

The project started in December 2016 when a message seeking collaborators was broadcast to a mailing list and two digital preservation specialists agreed to join the project to provide advice and guidance as needed. The work was then performed in the three phases illustrated in the chart below.



Scoping phase: Jan2017 - Feb2018: This work included completing the Scoping document and performing the tasks that the Scoping document identified as needing to be done before planning could start. Unfortunately, the collaborators had to leave the project halfway through this phase due to pressure of work. The final two months of this phase overlapped with the Planning phase.

Planning phase: Dec2017 - Feb2018: During this phase the Preservation Project Plan DESCRIPTION and CHART documents were produced.

Implementation Phase: Feb2018 - May2018: The tasks specified in the Preservation Project Plan DESCRIPTION and CHART documents were completed during this phase.

The work addressed ALL aspects of preserving a personal collection. Some of these aspects (such as the removal of a Document Management System (DMS) and the creation of a User Guide) may not normally be addressed within a digital preservation project, but they were considered important to the long term accessibility of the Collection. Decisions about what to do with particular files were taken with respect to what was practically feasible for the owner, and may not follow the approach that would be taken by digital preservation specialists and organisations. Such is the reality for personal digital collections.

**Project Start-up - Establishment of Collaborators**

I have little professional digital preservation knowledge and experience, so I decided to try and recruit some experts in the field to advise me in the work. My message to the JISC DIGITAL-PRESERVATION email list was answered by three individuals - two from Archives New Zealand and an Information Architect from Romania. The Information Architect only participated in the introductions and was not heard of after that - possibly due to upheavals in Romania at that time. My son, Matt Fox-Wilson, who is the potential future owner of the PAWDOC Collection should a permanent repository not be found, also joined the group.

These inputs to the work proved invaluable. They included specific knowledge such as how to make best use of the DROID file profiling tool [The National Archives, 2018], what were possible repository applications, and why specific files wouldn't open. It also included general advice on how to proceed, and feedback on documents. Given that all these individuals had full time jobs, I deliberately tried to minimise the amount of time they would have to spend on the project - though, in retrospect, the introductory conference call in which I introduced them to the Index and Document Management software was far too long and detailed. It would have been far better to keep things briefer and to respond to any requests for further detail.

**Scoping Phase**

The first version of the Scoping document was produced in January 2017 and was discussed on the Collaborator call on 02Apr2017. Section 9 (List all the activities that you will need to do BEFORE you are in a position to create a realistic plan for the digital preservation work that is needed) provided a useful set of discussion points from which action points were derived to drive the project forward; and it was from these action points that the following set of control documents emerged:

1. Alternative Document Management Systems
2. DROID Analysis
3. Files that won't open
4. Physical disks

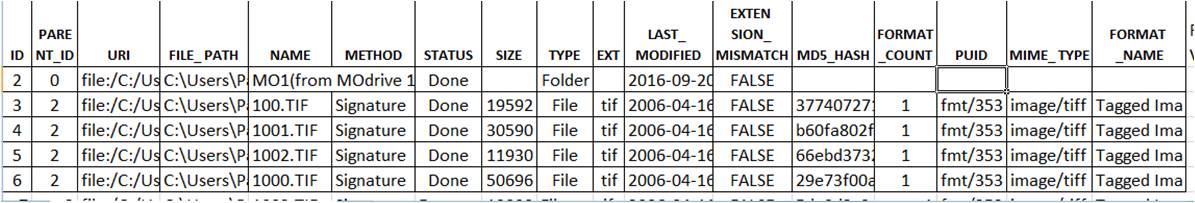
These addressed most of the preservation issues explored in the Scoping phase and which subsequently found their way into the tasks in the Preservation Project Plan. They are described in some detail below to give a flavour of the range of issues that are likely to be encountered in such an exercise.

1. **Alternative Document Management Systems**

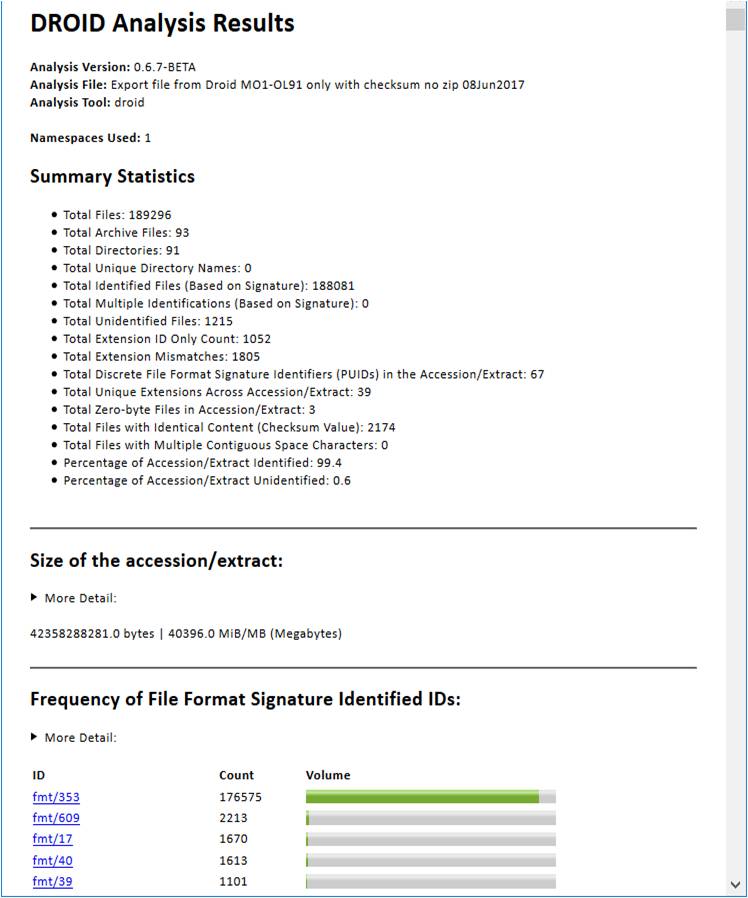
The Collection's electronic files were managed by the Fish Document Management System (DMS) and associated MS SQL database, which had been in place since 1996. Although the DMS effectively managed file storage and protected files against unintended change, movement or deletion, it was a complex product intended for organisations, not individuals; and it was expensive to upgrade and to obtain support for platform changes and configuration issues. Furthermore its future was uncertain and there were no plans to develop it further. These were documented as major risks to the Collection in the draft Scoping document. Therefore, one of the Collaborators wrote a draft 'Alternative Document Management Systems' document in which several preservation platforms were identified and described. I subsequently researched each of these on the net, but concluded that all were too complex and/or expensive for this purpose. At this point we started to consider the option of exporting the files from the DMS and storing them in Windows folders with Reference Numbers for the folder names. Eventually this was the solution that was decided upon and, from June onwards, I engaged in discussions with the supplier about using one of its utilities to export the files out of the DMS, and about the structure of the file names to be given to the exported files.

**B. DROID Analysis**

The DROID tool [The National Archives, 2018] is a free software application which analyses files and reports on their formats. It operates in conjunction with the National Archives PRONOM database of file formats and their associated PRONOM Unique Identifier (PUID). DROID was used to analyse the PAWDOC Collection's electronic files - all with cryptic file names allocated by the DMS and held in a specific Windows directory. DROID identified over 189,000 files (a very large number because the DMS held multi-page scanned documents as a series of separate TIF files, one for each page); and listed each one in an Excel file together with its pathname, size, file format, PUID, and a checksum as shown in the fragment below (some cell entries are truncated for the purpose of this presentation).

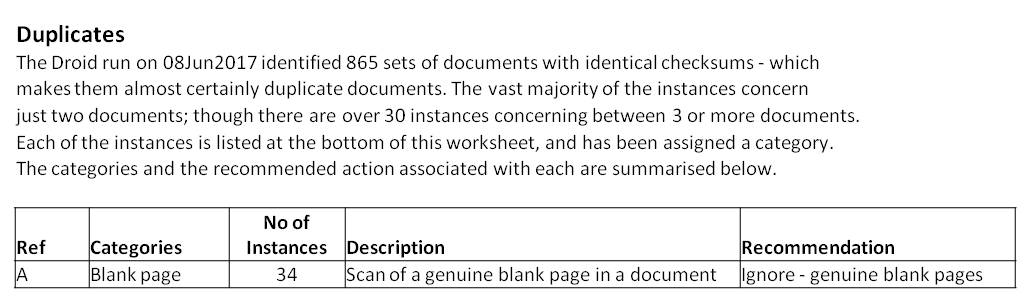


The checksum enables duplicate files to be identified and these were listed, along with a large amount of other detailed, very useful, information, in a summary report produced by one of the Collaborators using a DROID Analysis Tool [droid-siegfried-sqlite-analysis-engine, 2018] as shown in the snapshot below.



All these outputs were used to construct a DROID Analysis spreadsheet for the project, separated into the following key worksheets: Duplicates, Zip Files, Zero-Byte Files, Files not identified by DROID, and File type analysis.

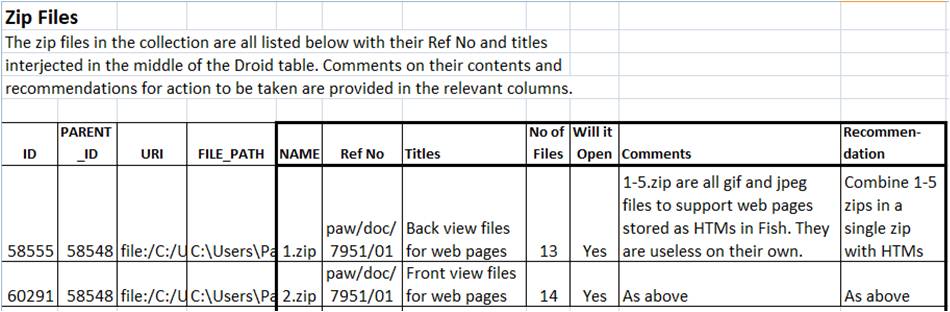
* Duplicates: An illustration of part of this spreadsheet is shown below.



865 instances were investigated of which about 35 involved 3 or more files:

* + About 35 instances were due to scans of blank pages resulting in exact duplications.
  + 380 instances concerned the same file being stored by the DMS with different file names, probably due to user error in the movement of files from Magneto-Optical disks into online storage in the 1990s.
  + about 320 instances of files being deliberately stored more than once in the DMS because they were specified in two or more index entries.
  + 50 instances of two or more identical versions of the same document (some undoubtedly created because the file was inadvertently opened in edit mode and then saved without any changes).
  + 12 instances seemed to be due to the DMS creating another file when a page image is rotated and the files being exactly the same.
  + In about 30 instances the duplicate files had different names but the DMS recognised only one of them.
  + 24 instances were due to human error in storing the same file twice in the DMS or of storing the wrong file in the DMS (all these 24 errors were rectified in the course of the Scoping investigation).

None of these instances, except for the 24 due to human error, affected the integrity of the Collection and therefore they were ignored in subsequent stages of the project.

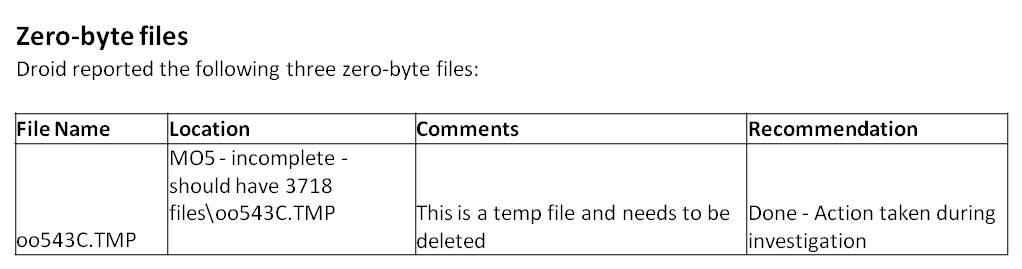
* Zip Files: An illustration of part of this spreadsheet is shown below.  
  

89 zip files were investigated:

* + 5 zips would need combining into a single zip.
  + 29 required no work and were retained in the Collection as zips in their current form.
  + 1 contained Visio files which would need converting to PDFs.
  + 22 were zips of emails in eml format which would require conversion to 22 single PDFs.
  + 4 contained GIF or HTML files which would need extracting/moving into 4 separate PDFs.
  + 1 contained a web site with linking errors which would need fixing before re-zipping the files.
  + 6 zips contained the contents of CDs which had been copied, but, for 5 of them, the relevant applications wouldn't open on the Windows 7 machine; and for the other (a PC magazine disk from 2001) there were too many files to look at as part of this project. For all of these it was decided to include a note in the Preservation Maintenance Plan (which would be finalised in the implementation stage of this project) to the effect that work on these items may need to be included in the next Preservation Maintenance project.
  + 2 contained Lotus Notes files which were converted into 2 PDF files in the course of the Scoping exercise (this involved downloading a trial version of Notes from the IBM web site, exploring what could be retrieved, and, having got to the point of understanding what could be done, it made sense to do the work there and then as opposed to marking it for inclusion in the downstream project plan).
  + The contents of 19 zips would need to be extracted and stored as base files.

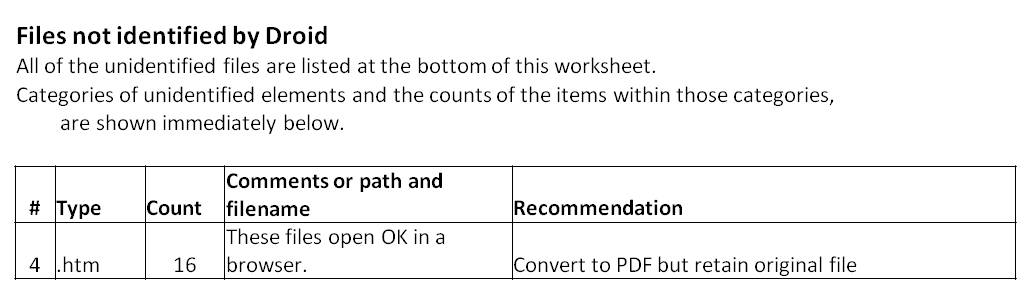
Overall, 29 of the 89 zips required no further work and the two containing Lotus Notes files were converted successfully during the Scoping phase, leaving 58 zips to be included in the downstream project plan.

* Zero byte files: An illustration of part of this spreadsheet is shown below.



Three such files were identified by DROID and all were dealt with during the Scoping phase. One was a temp file which was deleted; and two were empty files (one Access and one Word) which should not have been empty - notes were included in the relevant Index entries saying that this content had been lost

* Files not identified by DROID: An illustration of part of this spreadsheet is shown below.

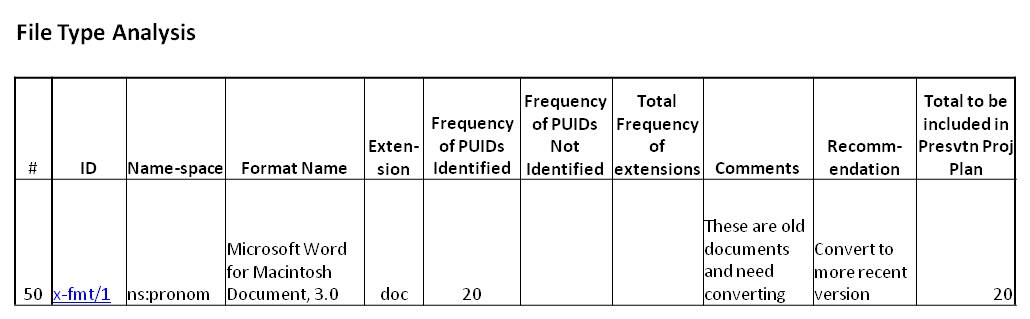


The underlying structure of 71 files was not identified by DROID:

* + 16 HTM files which, despite not being recognised by DROID, opened OK in a browser. They would require conversion to 16 PDFs.
  + 42 Filemaker12 files - These were in the latest Filemaker format which hadn't yet been included in the DROID PRONOM database, so they were ignored.
  + 5 files in earlier versions of Filemaker. However, for each of them, a copy in the latest version of Filemaker already existed in the Collection (every time a new version of Filemaker was installed, copies of all Filemaker files in the Collection were made in the new version). So, these files were ignored.
  + 4 SCM files. These were Lotus ScreenCam files and despite extensive searches on the net and attempted communication with the possible owners of the software, no means of converting these files was found. A note was included in the Preservation Maintenance Plan for these to be looked at again in the future.
  + 1 PDF file which the PDF application was unable to open because 'The base pages object is missing or invalid'. Since another (though earlier) version of this document was available in the Collection, the following note was added to the name of this file during the Scoping phase: 'Will not open - see earlier version of this brochure'.

It was concluded that, of all these 71 files, only the 16 HTMs would need to be included in the Project Plan. NB. DROID users are requested by The National Archives to report back details of files that DROID doesn't recognise so that they can update the PRONOM database [Clipsham, 2012].

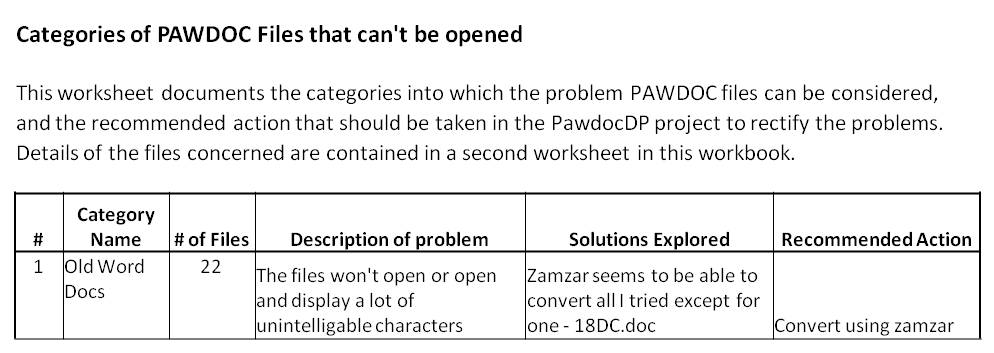
* File type analysis: An illustration of part of this spreadsheet is shown below.



Although most of the 189,000+ files that were categorised by DROID into file formats were not considered to be in any immediate danger of becoming obsolete, there were still a substantial number of files in older formats. Unfortunately, however, by the time the detailed investigation of these files was started, the Collaborators had had to leave the project due to pressure of work, and, consequently, decisions about exactly which of the older formats should be converted to newer formats had to be taken on the basis of information gleaned from the net. In the end, it was decided that about 750 files would need to be updated to later versions or to PDF: details are provided below with the National Archives PUID identifier included in brackets after each file format:

* + 1 WordPerfect for MS-DOS/Windows Document 5.1 file (x-fmt/394);
  + about 130 Word for Macintosh Document 3.0, 4.0, 5.0 files (x-fmt/1, x-fmt/64,x-fmt/65);
  + about 540 Word for Windows Document 1.0 & 2.0 files (fmt/37, fmt/38);
  + 3 Microsoft Word Document Template 97-2003 files (x-fmt/45);
  + about 20 Microsoft Excel 3.0 & 4.0 Worksheet files (fmt/56, fmt/57);
  + 24 PowerPoint for Macintosh 4.0 files (fmt/179);
  + 8 Microsoft PowerPoint Presentation 4.0 files (x-fmt/88);
  + 4 MIME Email 1.0 files (fmt/950);
  + 15 Rich Text Format 1.0-1.4, 1.5-1.6, 1.7, 1.9 files (fmt/45, fmt/50, fmt/52, fmt/355);
  + 11 Visio Generic files (fmt/442);
  + 2 Visio Drawing 2000 files (x-fmt/258) were converted to PDF during the Scoping phase.

**C. Files that won't open:** A previous check of the whole Collection in 2016 identified 220 files that would not open. Fifteen of these were dealt with in the DROID analysis described above. The remainder were investigated in this part of the Scoping work. An illustration of part of the spreadsheet used to manage this work is shown below.

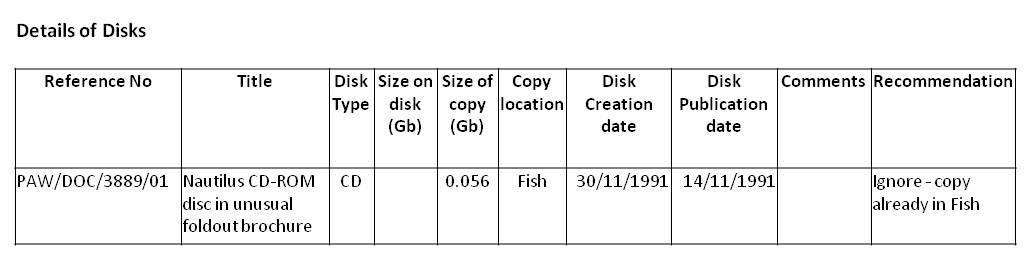


The major outcomes of this investigation are listed below.

* 22 of the files were old Word documents which would need converting by the Zamzar service [Zamzar, 2018] or by some other means.
* About 150 of the files were old PowerPoint documents which it was decided to update using the Zamzar service.
* 1 Microsoft Project file was investigated during the Scoping phase but no conversion service was found that could open it. The final person I asked for help from replied that "either it is corrupt or from a version of Project earlier than 4.x.". A note was put into the Preservation Maintenance Plan to the effect that another attempt to open this file should be undertaken in the next tranche of preservation work.
* 4 Lotus ScreenCam files were investigated during the Scoping Phase, but, despite making extensive trawls of the net, no conversion facility could be found. A note was included in section 3 the Preservation Maintenance Plan for these files to be looked at again in the future.
* 2 large MS Help files (which Microsoft stopped supporting in 2006) were converted to MS HTML Help (CHM) files during the Scoping phase by following the extensive guidance at a Help website [Help-Info.de, 2018], and by using the application 'HTML Help Workshop' which is available for download from the Microsoft web site [HTML Help Workshop, 2018]. The process involved the following rather complex stages: a) decompile the HLP file into its component parts, b) convert all the HLP components into HTML hhp files, c) set parameters in the hhp files, d) reconstruct the Table of Contents, e) re-insert spaces in headings, and f) compile the revised files into a single CHM file. Furthermore, the HTML Help File software was prone to crashing, so I quickly learnt to save regularly. Overall it took about 30-40 hours, but it was a success and the extensive contents of the two files were again fully accessible.
* 4 HTML files were dealt with during the Scoping phase: 1 was found to open up OK; the contents of 1 were retrieved and placed in another file; and two were found to have been incorrectly stored without any content so there was nothing that could be done apart from putting a 'Saved incorrectly - file has no content' in the titles of the files concerned.
* 2 PDF files were dealt with in the Scoping phase. One had its extension changed to JPG which resulted in it opening successfully. The other was found to be a damaged file. However, since there was another good version of the same file present in the Collection, no further action was specified.
* 1 TIF file wouldn't open within the DMS, however, in the course of investigating it in the Scoping phase, it was successfully exported from the DMS, converted to a PDF, and reinserted into the DMS in its new format.
* 1 XML file was dealt with during the scoping phase - a note was put into its file title to the effect that it can be viewed in the Notepad application.
* 3 BMP files were all dealt with in the Scoping phase: the extension of one was changed to TIF and the other two now seemed to open OK (perhaps something to do with the laptop being rebuilt after a crash in 2017).
* 1 Lotus 1-2-3 Worksheet 1.0 was successfully converted to Excel during the Scoping phase.
* 2 ITM files from the iThink application were emailed to ISEE Systems who converted them to a form which could be viewed on the downloadable isee player. Print Screens from the isee player were then copied into PDF files. All of this was done during the Scoping phase as, by the time I'd got to the point of understanding what could be done, I'd already just about finished the job.
* 3 Access files were dealt with during the Scoping phase by importing their data to Excel files. A fourth file was only 1Kb and wouldn't open via Excel. Since its title was 'Record locking file for Access Document Management database' it was concluded that it didn't contain any readable content and could be ignored (though it would still be present in the Collection if it was needed to open another file).

**D. Physical disks:**

16 physical disks in the Collection and 140 backup disks were reviewed in the Scoping phase and it was concluded that further work would be needed on 133 of them. An illustration of part of the spreadsheet used to manage this work is shown below.



The investigation included the following activities:

* ISO images of two CDs were created in order to explore the process of creating disk images in anticipation of applying it to several other disks. However, it was decided that zip files would suffice.
* It was decided that the contents of 127 of the disks would need to be copied, zipped and included in a special backup directory with the Collection's other files in the laptop.
* The contents of one 5.25 disk and two 3.5 disk were accessed and put on disk by the disk conversion service, Luxsoft [Luxsoft of Luxulyan, 2018]. This was done during the Scoping phase, and the files set aside for inclusion in the Collection after the export from the DMS.
* The video on one CD was converted to MP4 format from the original VHF tape by the video conversion service, Digital Converters [Digital Converters, 2018]. This, too, was done during the Scoping phase and the file set aside for inclusion in the Collection after the export from the DMS.

**Observations from the Scoping phase**

The following key points emerged concerning work that is done during the Scoping phase:

* It's worth putting a lot of effort into establishing a comprehensive record keeping tool (probably a spreadsheet) right at the beginning of the Scoping phase. With large volumes of files, it is easy to get confused or forget what you've done before or where you are up to. Each file, or instance of a problem, should be placed on its own separate row; and the related information to be recorded (i.e. the columns in a spreadsheet) should reflect the information you will need to manage and report the work. Columns should be included for all statistical information, so that any numbers you need are automatically calculated and visible (preferably above the column headings, not at the bottom of the spreadsheet), and don't have to be added up manually.
* The flexibility to be unstructured and to pursue unfamiliar avenues of investigation during the Scoping phase is a huge advantage. For example, the process of trying to track down organisations and people to convert specific file types can take varying lengths of time with no certainty of success.
* Full use of the internet should be made to identify advice, guidance and conversion services. For example, the approach taken to convert Lotus Notes files was found in a discussion forum.
* Knowledgeable collaborators are extremely useful to have on hand so that they can be asked for specific guidance about which file types to convert, how to convert them, and what to convert them to. It is probably not necessary to involve such people in the minutiae of the project unless they so wish it.
* For small numbers of relatively unusual files, by the time the conversion process has been discovered and tried out, you may have nearly completed the task. In that case, it is probably better to finish it off there and then within the Scoping phase rather than waiting to complete the work as part of the downstream project plan. Apart from anything else you may forget how to do it. This was definitely the best course of action for the Lotus Notes files and the MS Help files that were both dealt with in the Scoping phase of this project.
* If a solution can't be found in the Scoping phase and you don't want to spend any more time, don't be afraid to specify that those particular files or issues should be included in the Preservation Maintenance Plan. That way, the project doesn't get bogged down, and the problem areas don't get forgotten.

**Changes to the Scoping document template**

As a result of the experience gained in this project, the following changes have been made to the Scoping Template (the updated version of the template is embedded in this electronic file in Appendix B):

* The title of 3) was changed from ‘Why do you want to keep this data?’ To ‘Why do you want to keep this collection?’.
* Sections 5 (Why do you want to keep this data?) and 6 (For whom are you keeping it? etc.) were moved to after Section 2 (What are the main contents of the collection?) – this is a more logical flow of the contents.
* The title of section 6 was changed from ‘For whom are you keeping it? How are you going to test their expectations?’ to ‘For whom are you keeping it? What are their functionality, technology and any other requirements? How are you going to test their expectations?’.
* Section 7 (What risks do the different parts of the collection face?) and section 8 (What actions should be taken to mitigate the risks? Who is responsible for each action?) were combined so that actions are directly under the risks. The way it used to be, it was necessary to keep referring back to section 7 to see what the risk was you were talking about in Section 8.
* Section 12 (What are your expectations of quality) was removed - it is too broad a question at this early stage of the project. The achievement of quality standards for specific activities needs to be built into the project plan itself.

**Planning Phase**

Most of the Scoping work was completed by the end of 2017. The remaining elements involved working with the DMS supplier to establish exactly how the export from the DMS would work and what the format of the resulting file names should be. This activity continued right up to the start of the implementation phase. Consequently, the creation of the Project Plan DESCRIPTION document and associated Project Plan CHART was started early in January while the DMS export scoping work was continuing. The documents were finally completed when the DMS export arrangements were finalised on 19Feb2018.

**Observations from the Planning phase**

* The Principles, Assumptions, Constraints and Risks (PACRs) are particularly powerful prompts for driving out solutions (for example, assumptions about the longevity of certain file formats). So it's worth starting to think about them during the Scoping phase.
* Seeking input on the PACRs from experts in the field is well worth doing. Unfortunately, in this case, the Collaborators had to leave the project before the topic was addressed.
* If tasks which are still unquantifiable and which could adversely affect the project's timescales, still remain after the Scoping work has finished, consider specifying them as Risks to the project. They can then be given a mitigation whereby work on the tasks concerned is stopped after a certain number of days and an entry is made in the Preservation Maintenance Plan for the task to be looked at in the future.
* In order to define section 6 (Project Milestones and Deliverables) of the DESCRIPTION document, it is easier to do a quick draft of the Project Plan CHART first, and then to iteratively develop the two documents (section 6 and the Chart) in parallel.
* Unless there is great urgency, it is best to wait until ALL Scoping aspects have been completed before starting on the final version of the CHART and on sections 6 & 7 of the DESCRIPTION documents. Otherwise, the dates that have to be inserted into both documents will keep moving.

**Changes to the Plan DESCRIPTION and Plan CHART document templates**

As a result of the experience gained in this project, the following changes have been made to the Project Plan DESCRIPTION and CHART Templates (updated versions of these templates are embedded in this electronic file in Appendix B).

* A References section has been added to the end of the DESCRIPTION document.
* The option to include risk mitigation activities has been included in section 4.4 of the DESCRIPTION document.
* Various formatting changes have been made to the tables in the DESCRIPTION document.
* A ‘Date Done’ column has been added to the CHART document.

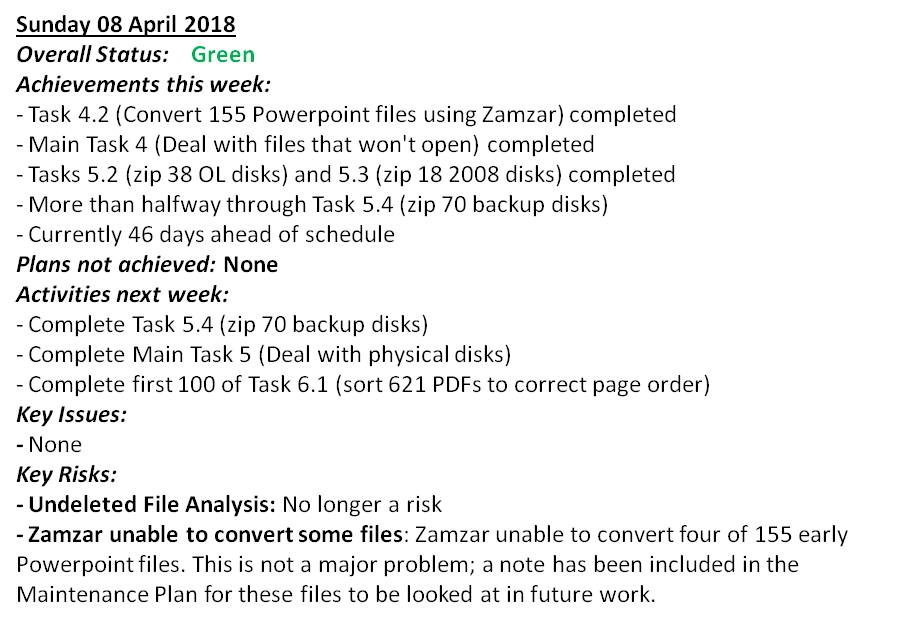
**Implementation Phase**

The implementation phase was started on 20Feb2018, and was completed on 03May2018 some 79 days ahead of schedule. Clearly, there was a serious overestimate of the time specified to do the work and there were two main reasons for this: first, without organisational constraints, I preferred to overestimate the time required and come in early rather than overshoot; secondly, I was motivated to spend a lot of time (i.e. all available hours) to make progress in the face of such large numbers of repetitive tasks - I couldn't face the prospect of having to plough through hundreds of files for many weeks - I just wanted to get it done as quickly as possible.

The Preservation Plan CHART remained on the desk throughout, and a completion date was written in as each task was finished as illustrated in the fragment below.



Progress was recorded in a weekly progress report compiled regularly on the same day each week as illustrated in the example below.



A few unforeseen problems arose in the course of the work and these were dealt with either by recording them as issues in the progress report and finding a solution; or by making an entry in the Baseline change log. No changes were recorded in the Project Plan DESCRIPTION amendment record.

A description of what occurred in each main task in the plan is provided below.

**Task 1 : Export files from the Document Management System (DMS) (**Planned: 26 days, Actual: 15) There was an initial hiccup with the export because the full scale test I performed had used the option 'export without delete' and worked perfectly. Unfortunately, when it came to do it for real, I needed to use the 'export with delete' option in order to be able to verify that all the files had been exported and to be able to check those that hadn't. This option didn't work correctly and it required a few days of iterations with the supplier before it did work. This provides a salutary reminder that unless all aspects have been tested you can never be absolutely sure that something will work. After the export had been completed, various checks were carried out and these confirmed that the export had worked as specified.

**Task 2: Adjust problem Zip files (**Planned: 16 days, Actual: 9).

Notable outcomes from the Zip file work were as follows:

* The Zamzar service [Zamzar, 2018] was used successfully to convert 9 Visio files to PDF and 17 PowerPoint files to PowerPoint 1997-2003.
* 22 zips contained eml files which opened OK in the MS Live Mail application when the Scoping work was being done. Unfortunately, Live Mail was not included in the reinstall of the Collection laptop after it crashed in November 2017, and it was no longer available on the MS web site. So, when it came to trying to convert these files I was not able to follow my original plan to open them in Live Mail and copy them to a PDF. I tried solutions suggested on the net including changing the extension to .mht and opening in a browser (it worked but the message header was not displayed), and installing the Outlook application (it wouldn't open the files). In the end, I installed the free Mozilla Thunderbird email application, opened each email in turn, took a Print Screen of it and pasted it into a PDF which accumulated all the emails from a single zip. It took quite a long time to do this to all 199 emails but at least I did get a result and the emails can be easily read in the22 PDF files that emerged from the process. The main disadvantages of such an approach are that a) its very time-consuming, b) sometimes the bottom of one page of text is duplicated at the top of the next pages, and c) a text search cannot be performed across print screen pages unless Optical Character Recognition is applied to each page. Interestingly, one of the Collaborators who reviewed this paper said that, if he had still been involved in the project when the eml problem surfaced, he would have suggested investigating Print-to-File from Thunderbird, which would have retained the page-breaks and would have been quicker.
* The need to consider doing preservation work in the future on 6 zip files was recorded in section 3 of the Preservation Maintenance Plan.
* Around 70 files were extracted from 15 zips and stored as files in their own right, and a further 140+ files from 7 zips were combined into 5 PDF and 2 Excel files (in Excel each file was stored as a separate worksheet).
* It was decided not to store the contents of 2 zips as base files as originally planned because of the quantity and similarity of the files contained within the zips.
* After the first few file conversions were completed, it was decided that it would not be useful to remove old versions to a separate directory and an entry was made in the Baseline Change Log specifying that both old versions and new versions would be retained in the same folder.

**Task 3: Adjust problem files identified by DROID) (**Planned: 30 days, Actual: 13).

Notable outcomes from the work done on problem files were as follows:

* It was decided not to convert 16 htm files to PDF because it was found that the converted files contained unformatted text embedded within html code - which was highly unreadable (this hadn't been tested during the Scoping phase).
* Zamzar [Zamzar, 2018] was used to convert 11 Visio files to PDF, though, in two cases, the converted files contained some obscured text, and some text moved to another line, respectively. These files were converted manually by opening them in a browser window, taking a screenshot, pasting it into PowerPoint, cropping it, saving it as a JPG and importing the JPG into the eCopy PDF Pro application to produce a PDF. In another case, Zamzar produced a blank page; however, from the file title, this was assumed to be a Visio Stencil configuration file with no readable content, and therefore not important to the Collection.
* 1 WordPerfect for MS-DOS/Windows Doc 5.1 file was converted to PDF.
* 20 Word for Macintosh 3.0 files were converted to DOCX - although only after eventually discovering that a lack of formatting in the documents could be resolved by doing a carriage return somewhere in the document.
* 101 Word for Macintosh 4.0 and 5.0 were converted to 78 PDF files and 23 DOCX files (the 23 files were retained in Word DOCX format rather than moving them into PDFs because the formatting and pagination that came through when the files were opened in Word 2007 was not quite right, and leaving them in native Word format would enable a reader to rectify such problems).
* 466 Word for Windows Document 1.0 and 2.0 files were converted to 409 PDFs and 57 DOCX files (the DOCX format was selected for the same reason as described for the Word for Macintosh 4.0 & 5.0 documents above).
* 4 Word for Windows 1.0/2.0 files couldn't be opened because memory problems prevented the file opening fully. A note recording that future preservation work will be required on these files, was included in the Preservation Maintenance Plan due to be produced in Task 8.
* 1 file opened successfully in the browser despite DROID identifying it as a Word for Windows 2.0 file, so this was retained without any change to its .htm extension.
* 3 Word Document template 97-2003 files were converted to 2 PDF and 1 DOCM files.
* About 20 Excel 3.0 and 4.0 files were converted to XLSX files.
* Zamzar [Zamzar, 2018] was used to convert 4 PowerPoint for Macintosh 4.0 and PowerPoint Presentation 4.0 files to PPT and then they were changed to PPTX locally (and three of them were also converted to PDF).
* 4 MIME Email 1.0 files were converted to 1 DOCX and 3 PDF files.
* 15 Rich Text Format 1.0-1.4, 1.5-1.6, 1.7 & 1.9 files were converted to 15 PDF files.

**Task 4: Deal with Files That Won't Open (**Planned: 11 days, Actual: 5).

Notable outcomes from the work done on these files were as follows:

* 10 were Casewise files (which had been saved in the DMS as Word documents) for which a conversion facility could not be found. Their extensions were changed to either .dat or .ind (which is what they were before being stored in the DMS), and a note was included in section 3 of the Preservation Maintenance Plan to the effect that these files will need preservation work in the future.
* 1 file was a Lotus Organiser v2 file for which a conversion facility could not be found. The extension was changed to OR2 (as it was before being stored in the DMS) and the need to do future work on these files was included in section 3 of the Preservation Maintenance Plan.
* 1 Word (Generic) 6.0-2003 files was converted by opening it in Notepad, cleaning up spurious characters, saving it as a DOC file, and then converting it to DOCX.
* 3 Word for Macintosh Document 4.0 & Word (Generic) 6.0-2003 files were converted by Zamzar [Zamzar, 2018] to DOCX files.
* About 150 PowerPoint for Macintosh (4.0) & PowerPoint Presentation 4.0 & PowerPoint Presentation 97-2003 files were converted by Zamzar to PowerPoint 2003-7 files (Zamzar would not convert to PPTX format which would have been preferred). However, in 4 cases some minor editing was required to fix the appearance of the converted slides.
* Zamzar was unable to convert 2 PowerPoint for Macintosh (4.0) and 2 PowerPoint Presentation 4 files so a note was placed in section 3 of the Preservation Maintenance Plan to the effect that future work may be required on these files.

**Task 5: Deal with Physical Disks (**Planned: 24 days, Actual: 7).

125 disks were opened successfully and copied to the laptop, despite some being over 20 years old. Some of them didn't open straight away but just continued whirring. Cleaning with a cloth didn't seem to help. However, what did work was selecting ‘Computer’ on the left side of the Windows Explorer window to display the laptop’s own drive on the right side of the window together with any external disks that are present. For some reason which I don't understand, disks which kept on whirring without seeming to be recognised did appear on the right side of this window and I was able to select and copy them to a separate Backup directory for the Collection. With more knowledge and resources, perhaps I would have investigated these strange symptoms further to ensure there were no underlying problems with the data. However, I just wanted to get the job finished, and, given that there were no obvious problems with the files, I just carried on.

Two of the disks (both CDs) failed to open despite all my efforts. One contained the software for an old version of the DMS application used by the Collection so the content is of little importance - its mere presence in the Collection tells its own story. The second disk contained backup files associated with the DMS and some of the files it held, and these will probably never need to be accessed. If any of the contents of these disks had been important to the collection, I would have sent them to the Luxsoft service [Luxsoft of Luxulyan, 2018] to try and have their contents retrieved. However, that course of action did not seem necessary.

**Task 6: Deal with Double Unsorted files (**Planned: 28 days, Actual: 12).

The pages in 881 files were re-ordered to rectify the page order produced by scanning all front sides first and then turning over the stack of pages to scan the reverse sides at a time in the 1990s when I didn’t have a double sided scanner. The re-ordering was done by first converting each file from a multi-page TIF file to a PDF, and then opening the PDF to re-order the pages.

My PDF application – eCopy PDF Pro – had two functions which made this task a whole lot easier: first, the function to have eCopy convert a file to PDF is available in the menu brought up by right clicking on any file, thereby automatically bringing up a Save As dialogue box with a suggested file title based on the title of the original file; and which, after the conversion is complete, then automatically displays the newly created file. This is a relatively quick and easy process.

Second, eCopy has a function whereby thumbnails of all the pages in a document can be displayed on the screen and each page can be dragged and dropped to a new position. I soon worked out that the front-sides-then-reverse-sides scan produces a standard order in which the last page in the file is actually page 2 of the document; and that if you drag that page to be the second page in the document, then the new last page will actually be page 4 of the document and can be dragged to just before the 4th page in the document. In effect, to reorder simply means progressively dragging the last page to before page 2 and then before page 4 and then before page 6 etc until the end of the file is reached. Both these functions (to be able to click on a file title to get it converted, and to drag and drop pages around a screenfull of thumbnails) are well worth looking for in a PDF application.

**Task 7: Revise the backup and DR arrangements (**Planned: 7 days, Actual: 3).

The Collection already had a full range of backup arrangements in place - cloud, external hard disk, other household laptop, UK remote hard disk, and New Zealand remote memory stick. A new dedicated external hard disk was purchased just before the start of the Implementation phase to ensure no losses during the huge changes that the preservation project was making to the Collection. This task documented the backup arrangements and devised a way of including them in the User Guide (see Task 9) such that a multi-year completion form appeared on the back of the Guide.

**Task 8: Create Preservation Maintenance Plan (**Planned: 7 days, Actual: 3).

It was relatively quick and easy to complete the Preservation Maintenance Plan. The following sections needed filling in:

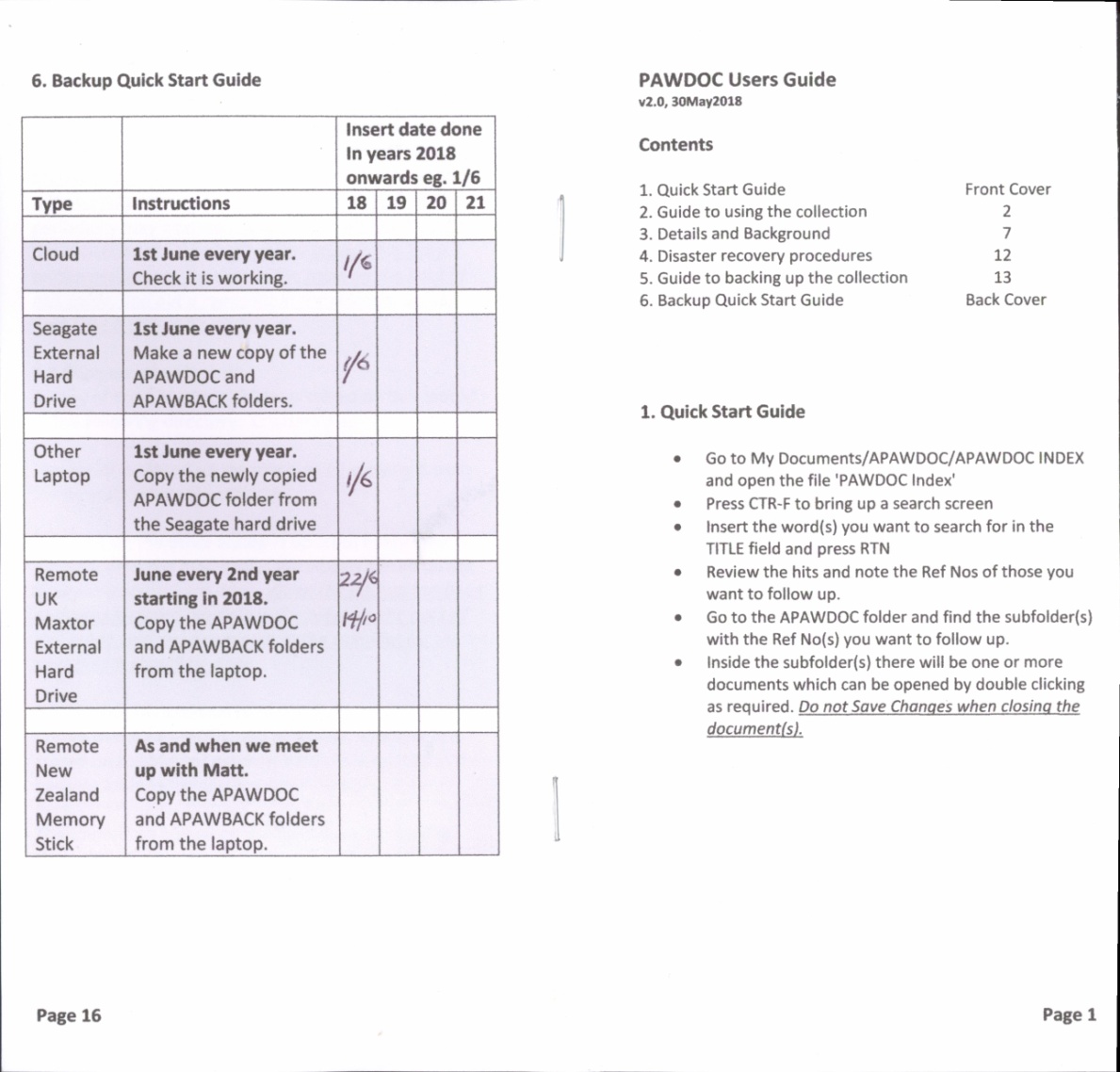
* **Schedule:** the date of the next Maintenance exercise was set at 1st September 2021.
* **2. Previous preservation actions taken:** This required a summary of what had been achieved in the current exercise and was quite difficult to derive - particularly because some items appeared in more than one category of work. The structure of the control documents that are used within a preservation exercise should be devised to facilitate the completion of this section.
* **3. Possible future preservation issues:** This section had already been completed in the course of the preservation exercise. It recorded issues with the following:
  + 6 zip files;
  + a pair of executable files;
  + memory problems with four Word documents;
  + four PowerPoint files that Zamzar couldn't convert;
  + 1 MS project file, 4 Lotus Screencam files, and 10 Casewise files, none of which could be converted;
  + 34 A5 documents with disordered pages in multi-page TIF files;
  + Probable non-conformance to the PDF-A-1b standard.

**Changes to the Preservation Maintenance Plan template**

As a result of the experience gained in this project, the following changes have been made to the Preservation Maintenance Plan Template (updated versions of these templates are embedded in this electronic file in Appendix B):

* **Section 2 (Previous preservation actions taken):** The suggested text starting, 'The formats...' has been changed to:   
  'The formats of all @@@ files <Insert the approximate number of files in the collection.> comprising the digital collection had been reviewed and the following conversions performed to ensure the files concerned remain readable in the future. All files that were converted kept the same file title as the original but had 'UPDATED' and the date of the conversion inserted into their titles. The original versions were retained but had 'ORIGINAL' inserted into their file titles. <Remove this last sentence if it does not apply.>'.
* **Section 4.** **How to implement this Maintenance Plan:** The paragraph in the middle of the list:' If required, create a Digital Preservation Plan DESCRIPTION and CHART at this point' has been changed to: 'If required, create a SCOPING document, or a Digital Preservation Plan DESCRIPTION and CHART at this point.'
* **Throughout:** A variety of other formatting and minor wording changes have been throughout the document.

**Task 9: Produce User Guide (**Planned: 5 days, Actual: 1).  
It had been determined that two copies of the Guide would be needed - one for inclusion in one of the two hardcopy archive boxes, and one for the box of backup disks. The Archive boxes have flexible plastic CD wallets stapled to the inside of the flaps and one of these was empty. So, the User Guide was designed as a 16 page A5 bookfold document with sufficient side margins to enable it to be cut down to fit into the CD wallet. The front page was devoted to a short contents list and to a 6 point 'Quick Start Guide'. The back cover was devoted to a 'Backup Quick Start Guide' in the form of a table enabling completion to be signified by writing the date in the appropriate cell. A scan of the front and back pages is shown below.



The other contents of the User Guide were:

2. Guide to using the collection

3. Details and background

4. Disaster recovery procedures

5. Guide to backing up the Collection

**Task 10: Close down PawdocDP project (**Planned: 4 days, Actual: 3).

Final checks were made on the newly created folders and on the Index, and a few adjustments were made. The directory containing all the working documents associated with the project was looked through and key documents were stored within the Collection. All other documents were deleted so that the directory was empty at the end of the process. The Plan also required that new remote backups be created at this stage. However, the newly created backup documentation specified that this task was to be performed after 1st June, so an entry to that effect was made in the Baseline Change Log.

**Observations from the Implementation phase**

* Problems were dealt with either by recording them as issues in the Progress Report and finding a solution; or by making an entry in the Baseline change log; or by making an entry in Section 3 of Preservation Maintenance Plan so they could be investigated at a later date.
* The Weekly Progress Report was an effective tool for encouraging issues to be surfaced and dealt with; and for providing motivation to complete chunks of work.
* Some files appeared in both the DROID and the 'Files that Won't Open' categories of investigation. This was confusing. Furthermore, it was difficult to derive the summary information about preservation actions that had been taken for section 2 of the Preservation Maintenance Plan, from the control spreadsheets. It would have been well worth getting the structure of the spreadsheets right in the Scoping Phase.
* Although ORIGINAL versions of files that have been UPDATED are being kept in the main Collection, it is not clear if this is the right long-term approach.
* It's better to overestimate and come in early than underestimate and come in late.
* Despite putting a lot of effort into the Scoping phase, and in testing the DMS export utility, some unknowns were still encountered in implementation. The more investigation that can be done in the Scoping phase the better.
* If the Scoping phase takes a long time, things may change after plans are made, for example, it was planned to use Live Mail to open eml mail messages; but a system crash occurred after that decision was made and Live Mail was not available in the rebuilt system.
* The Zamzar service [Zamzar, 2018] was used very successfully to convert about 170 files, and only failed in 4 cases.
* Converting hundreds of files is undoubtedly a slog. Motivational aids such as spreading conversions across several tasks, or setting intermediate goals for tasks, are worth exploring.
* Combining the user guide and the backup documentation, and having quick guides to each, one on the front and back covers respectively, seems to work well.
* The ability to save UPDATED versions of MS Office files in DOCX, XLSX and PPTX format makes it easy to distinguish those files from older versions with DOC, XLS and PPT extensions.

**Conclusions**

The project was a success - all major objectives were achieved and the Collection now has a Preservation Maintenance Plan in place.

The Scoping phase was integral to the project's success. The more that can be discovered, tested and verified at that stage the better.

It is worth taking the time early in the Scoping phase to construct control spreadsheets which will support both the planning activity and the subsequent implementation actions; and which will facilitate progress reporting and the provision of information to Section 2 of the Preservation Maintenance Plan.

The Templates were very useful in getting the Scoping and Planning phases off to quick starts, and in providing direction for the work.

Although a Preservation Maintenance Plan has been constructed, the validity of that particular Template still needs testing in a follow-on preservation exercise.

Some preservation decisions will depend on the Collection objectives. For example, in this case, the objective is to provide users with readable content and to provide them with an awareness of the activities the original Owner was performing. It is NOT intended to provide users with a simulation of the experiences of the Owner. Hence, it was decided not to undertake preservation activities on some files on the basis that their mere presence in the Collection was sufficient. To have tried to resurrect some of those files and to get them to actually work would have probably involved huge amounts of work. It is clearly very important that such objectives are fully understood before setting out on a Preservation Planning exercise.

This project indicates that owners of personal collections with little knowledge of, or access to, professional preservation tools and techniques, can still make a success of addressing preservation issues. However, they would be well advised to seek guidance from Preservation professionals to help them in the work.

**Acknowledgements**

I would like to thank the following people for the help they have provided in the course of this work: Ross Spencer of Artefactual (ex Archives New Zealand) for his help in the Scoping phase of the project , in the use of Droid and the analysis of its results, and for providing extensive feedback on this paper; Jan Hutar of Archives New Zealand for his help in the Scoping phase and in the investigation into alternatives to the DMS; Matt Fox-Wilson for his help and support throughout the project; and Bridget Bateman of m-Hance for her generous and very helpful support over the last seven years of my use of the Fish DMS and for facilitating the export of files out of Fish into windows folders.

**References**

Clipsam, David, An introduction to the PRONOM contribution model and the Signature Developer role, National Archives website visited on May 18, 2018.<http://blog.nationalarchives.gov.uk/blog/an-introduction-to-the-pronom-contribution-model-and-the-signature-developer-role/>

Digital Converters, Website accessed on May 11, 2018. <https://www.digitalconverters.co.uk>

Droid-siegfried-sqlite-analysis-engine, Website accessed on May 18, 2018 <https://github.com/exponential-decay/droid-siegfried-sqlite-analysis-engine>

Help-Info.de. Website accessed on May 11, 2018. <http://www.help-info.de/en/Help_Info_WinHelp/hw_converting.htm>

HTML Help Workshop, Download available from the Microsoft website accessed on 29May2018. <https://www.microsoft.com/en-gb/download/details.aspx?id=21138>

Luxsoft of Luxulyan. Website accessed on May 11, 2018. <http://www.luxsoft.demon.co.uk/>

The National Archives. File Profiling Tool (DROID). National Archives website accessed May 11, 2018. <http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>

Wilson, P. **“**20years in the life of a long-term empirical personal electronic filing study.” Behaviour & Information Technology20, no.5, 2001: 395-409. <https://www.tandfonline.com/doi/ref/10.1080/014492901317188499?scroll=top>

Wilson, P., "Preservation Planning for Personal Digital Collections", April, 2016. Digital Preservation Coalition Case Note. <https://www.dpconline.org/blog/preservation-planning-for-personal-digital-collections-by-paul-wilson>

Zamzar, Website accessed on May 11, 2018. <https://www.zamzar.com/>

**Appendix A - Documents produced for the Project**

The following documents were produced for the project and are available on the pwofc.com website in a post dated 07March2019 within the topic ‘Preservation Planning’:

* PawdocDP SCOPING Document - v1.0, 05Sep2017, 03-05-2018.docx
* PawdocDP Preservation Project Plan DESCRIPTION, v1.0 - 19Feb2018.doc
* Completed PawdocDP project Plan CHART with finishing dates for each task, 03-05-2018.pdf
* PAWDOC Preservation MAINTENANCE PLAN - v1.1, 28May2018.docx

**Appendix B – New versions of the Preservation Planning Documentation Templates**

The updated templates below can be used as the basis of your preservation planning documentation. They include removable guidance on how to complete them. Anyone who uses the templates is encouraged to contact the author with feedback on their experiences so that the templates can be improved.

* Scoping Document Template 
* Preservation Project Plan Description Template 
* Preservation Project Plan Chart Template 
* Preservation Maintenance Plan Template 

The templates have also been published on the web site of the Digital Preservation Coalition (DPC) at the following address: <https://www.dpconline.org/blog/1095-preservation-planning-for-personal-digital-collections-by-paul-wilson>

This paper is shared under the following Creative Commons Licence:

