Endangered treasures in our library basements: Securing long-term access to content on CD ROM

Yvonne Friese; Digital Archivist; Kiel, Germany

Abstract

This paper provides a hands-on report of the ingest workflow of hand-held media in the library stacks of the Leibniz Information Centre for Economics. It describes inspection, selection, data carrier migration, the securing of the original structure, and ingest into the Rosetta Archive as they have been practiced by the library since 2010. The description includes some information about the content to be archived, the file formats, the quality of the data and related preservation issues, the tools used, the lessons learned and future work to improve the workflow. Furthermore, a brief overview of current best practice workflows is given.

Inspection of the library stacks

An analysis of our library catalogue has shown that our library's stack rooms contain almost 5,200 independent bibliographic items on hand-held media. In addition, about 2,600 printed materials have external data carriers attached. The mixture is heterogeneous; mostly CD ROM (2,200), floppy disks (290, of different kinds, mostly 3,5" and 5,24"), CD (42, e. g. audiobooks) and DVDs (91). Besides, the no. of items containing external data carriers is growing, as many new acquisitions contain CD ROM, DVD ROM or even an USB Stick.

The data carriers – and the file formats some of them include – are at risk to become obsolete in the near future, some of them are already difficult to access. Waiting will only increase the danger of losing content as the likelihood of obsolescence grows every year.

Best Practice

Fortunately, other public memory institutions already have developed workflows for similar tasks and have shared the findings and lessons learned with the community.

The British Library has stabilized at least 16 terabytes of data from hand-held media [1]; to achieve this, the data carrier migration was partly automated by the use of disk copying robots. However, due to the heterogeneity of the hand-held media, personnel involvement for exceptions and defect disks was necessary, as was the manual extraction of some important preservation metadata. They were able to stabilize 1,050 disks per month [1].

The NLNZ has needed 52 hours personnel time to archive 188 floppy disks (3,5") which correlates to almost 30 minutes per item [4].

The generation of metadata proved to be the most time-consuming factor of the workflow at the British Library [1] as well as at the NLNZ [4]. Furthermore, the NLNZ [4] stated that the preservation of the original folder and file structure of the data carrier can be easily lost during workflow steps like repairing the file extensions, file names or the normalization of file formats which are not up to date. They recommended preserving the file structure at a very early stage during the archiving workflow [4]. The National Library of Australia (NLA) also made the preservation of the original structure, including all hidden or deleted files part of the workflow [2].

In 2008, the Indiana University Library created a collection of almost 3,000 CD ROMs and was able to provide web-based access. They made preferential use of open-source software, the CD ROM content was made available via ISO images and the METS metadata was put on an Andrew File System (AFS) to encourage other libraries to share it [6]. Some of the material still derived from the MS-DOS generation or Windows 3.1. As the user should be able to just browse to the content and get it in formats that are easy to read and access, a lot of migration and emulation work was needed to fulfill these needs.

In 2013, there have been more workflows in progress concerning hand-held media, such as a workflow model for Audio CD Preservation, which has different needs in comparison to CD ROM and even less experience within the community yet [5] and some even more spectacular recovery projects like the one with floppy disks 5,24" at the University of York [3].

Data carrier Migration and Ingest to Rosetta

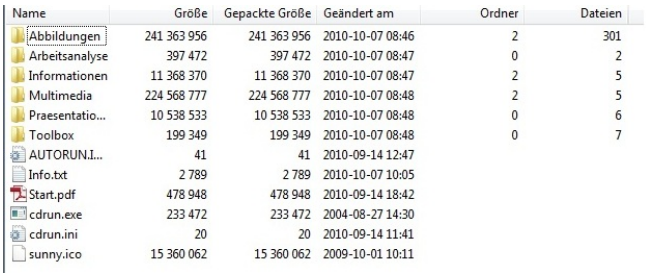
Unlike the British Library, we only have about 10,000 data carriers in our stacks. We decided against using a copying robot, because the experience of the British Library has shown that the data carrier migration with the help of robots still is very time-consuming and the implementation of this workflow is, too. The selection what to archive first is left to our users. Whenever a medium is requested, it is archived afterwards. Currently, about 50 units from external data carriers per month are demanded by our users. We intent to establish a second workflow for items newly acquired by our library. Following these two workflows, there will be a blind spot for material in the stacks which is not demanded by our users. It has not been decided yet whether to archive this material or not. This is a matter of personnel resources and the priority has to be the material that is often demanded by our users and the newly acquired material.

Dappert [1] has stated in her paper that the generation of metadata proved to be especially time-consuming. Therefore, some of the metadata generation within our workflow has been automated.

During the ingest process, there are seven Dublin Core metadata fields to be filled in manually:

1. dc.title (mandatory)
2. dc.identifier (mandatory)
3. dc.creator
4. dc.date (year of publishing)
5. dc.type (attachment or independent bibliographic unit)
6. dc.extent (No. of data carriers per unit)
7. dc.medium (CD, CD ROM, DVD,…)

To facilitate the upload into the archive, the content of the data carrier is compressed into a .zip file. The system decompresses the .zip file during ingest. In this way, the information about the original structure is saved in the METS file attached to the archival package. Furthermore, DROID and JHOVE extract technical metadata, a virus check is performed and checksums are generated for each file. In addition, the enrichment of the full set of bibliographical metadata is done automatically via a SRU interface from the union catalogue.



Name	Größe	Gepackte Größe	Geändert am	Ordner	Dateien
Abbildungen	241 363 956	241 363 956	2010-10-07 08:46	2	301
Arbeitsanalyse	397 472	397 472	2010-10-07 08:47	0	2
Informationen	11 368 370	11 368 370	2010-10-07 08:47	2	5
Multimedia	224 568 777	224 568 777	2010-10-07 08:48	2	5
Praesentatio...	10 538 533	10 538 533	2010-10-07 08:48	0	6
Toolbox	199 349	199 349	2010-10-07 08:48	0	7
AUTORUN.I...	41	41	2010-09-14 12:47		
Info.txt	2 789	2 789	2010-10-07 10:05		
Start.pdf	478 948	478 948	2010-09-14 18:42		
cdrun.exe	233 472	233 472	2004-08-27 14:30		
cdrun.ini	20	20	2010-09-14 11:41		
sunny.ico	15 360 062	15 360 062	2009-10-01 10:11		

***Figure 1**. Typical CD ROM in our stack*

In this way, each file can be retrieved easily, which facilitates the risk management based e. g. on file formats and the preservation planning and actions like format migration.

Initially, it was considered to transfer to the ISO file format (ISO 9660) as it is the standard data format for CD ROM contents. As far as the mere bit-preservation is concerned, this would have been a fair decision, as the ISO image is the "bit faithful" copy of a CD ROM [6] and would be very close to the original data carrier. Unfortunately, tests have shown that the ISO format is not handled well by our standard Ingest Workflow and the Preservation Planning Module used by our Digital Archive. The virus check, the technical extraction, the file identification and validation would be done on the ISO file only. The risk management would consider only the ISO file and does not look into the ISO file to evaluate the file formats hidden in the ISO package. As the file formats of our collection have turned out to be numerous, we want to collect as much information about it as possible and build a reliable risk management and preservation planning based on this information.

The files on a typical CD ROM often depend on and link to each other, as e. g. a bundle of HTML files or style sheets related to some XML files. Some of the HTML files cannot be read or do not make sense as stand-alone objects.

So it is crucial to archive the material of one CD ROM in one unit in order not to lose the coherence and additional information. One intellectual entity in our Rosetta archive correlates to one bibliographic item which consists of at least one CD, sometimes more than one, which is then noted to the dc.extent field.

File Formats on CD ROM

For a sample, 136 units consisting of or containing external data carriers like CD ROM were analysed in detail, with the help of DROID and a self-developed program which extracts and counts the file extensions.

On average, the size of the CDs is 240 MB, but the heterogeneity is big, as some CD ROMs contain less than 1 MB data material and some DVD-ROMs more than 4 GB.

There are 26,101 files on the 136 units overall, so the average contains 190 files. 177 different file formats have been found in the sample, 20 of them occur more than a 100 times, 54 at least 10 times and the remaining 123 file formats less than 10 times.

Table 1. The 19 most common file formats used on the CD ROMs

range	No. of files	Extension	Format
1	11,281	htm	Hypertext Markup Language
2	3,972	gif	Graphics Interchange Format
3	3,302	pdf	Portable Document Format
4	2,393	jpg	JPEG File Interchange Format
5	757	xls	Microsoft Excel Worksheet
6	483	doc	Microsoft Word
7	340	js	JavaScript File
8	274	txt	Text-File
9	241	cab	Windows Cabinet File, cannot always be detected by DROID
10	205	mtw	OLE2 Compound Document Format
11	181	class	Java Compiled Object Code
12	171	dll	Windows Portable Executable
13	167	css	Cascarding Style Sheet
14	163	xml	Extensible Markup Language
15	144	exe	Windows New Executable / Dos Executable
16	144	rtf	Rich Text Format
17	135	swf	Adobe Flash, Macromedia Flash
18	132	xpt	Statistical Analysis System Catalogue
19	102	ppt	Microsoft Power Point

DROID, however, is not without fail. As Wood [6] stated, DROID has given a false positive for a Microsoft Word Document with the file extension ".doc", but the files contain some binary data. When the Indiana University Library migrated the MS Office files to OpenOffice files, this had led to data loss for these files.

There are some files which cannot be identified by DROID, because the Pronom Library cannot possibly contain every file format ever having been in use.

Also among the 19 most frequent file formats in the sample, DROID cannot always detect the file format with certainty, even if the file extension is known to its library in principle, as e. g. some files with the extension ".cab"(Windows Cabinet File). DROID often determines the file format according to the internal Signature instead of by the file extension, which has proven to be the better strategy, as the extension can be misleading. An example of some XML files with a ".pdf"-extension will be described later in this chapter.

We have begun to build a criteria catalogue for migration for the most commonly used file formats. Table 2 exemplarily shows the preservation strategy for the six biggest groups of file types.

Group 1: Markup Languages
Examples: HTML, XML
Archived Format: original
Known Issues: The quality of the HTML files is difficult to measure. JHOVE considers a 100% of the tested sample to be invalid, which is quite typical for HTML files. As the browsers are flexible and tolerant, this should not be a big problem. Nevertheless, some of the HTML files are older than ten years and just do not look fancy any more in modern browsers, some can no longer be opened at all. There has to be an automated workflow to test if the HTML file is still accessible or not, and as some of the HTML files have dependencies to other files and can only be displayed properly in their original context, this is not a trivial task.

Group 2: Adobe Products
Examples: pdf, pdb, pfm
Archived Format: original
Known Issues: Re-use of already established PDF-repair Plug-In based on iText.

The quality of the PDF files is pretty good, only 7 % are considered to be malformed or at least invalid by JHOVE. These findings are much better than we are used to in other collections, where up to 50% of the files have issues with validation or well-formedness. Tests have shown that some of the files that purport to be a PDF are in fact no PDF files at all. While performing a test to find any encrypted PDF files, it became clear that some of the PDF files lack a PDF header and contain a xml header instead. While aiming to save a PDF file from a website, the site returned a 404 HTML page which the user did not realize and therefore saved the contents like he would a PDF file. The file looks like a PDF but must be treated like an xml or HTML file and lacks the kind of information is purports to contain. In fact, if there will ever be such a thing as a garbage selection, these kinds of files would be first in.

Group 3: Graphic File Formats
Examples. GIF, jpeg, png, tif
Archived Format: original
Known Issues: JHOVE: Test well-formedness + validity, errors fixed by Plug-In yet to be implemented (exp. for JPEGs)

Group 4: Simple text formats
Examples: txt, rtf,
Archived Format: original
Known Issues: none

Group 5: MS Office products
Examples: MS Word, Excel, Powerpoint
Archived Format: migration to OpenOffice, PDF (to be implemented & tested)
Known Issues: JHOVE struggles to check for well-formedness & validity

At the Indiana University Library the MS Word and Powerpoint documents were converted to PDF [6]. The examination of our files has shown so far that at least for the MS Word documents such a migration action is preferable for us as well. As for Powerpoint documents, there might be a loss of functionality, as the PDF Version of the Powerpoint Presentation lacks animations. If this would be considered to be among the significant properties for future users, another solution for this material has to be found.

As for the quality check for MS Word files JHOVE has not proven to be useful as it considers it in principle to be a "well-formed and valid bytestream". DROID might give a hint, because with difficult MS Word files it is often unable to detect the exact file format by signature, but makes a guess because of the extension and gives up to eight solutions about the possible file format. The MIME type was always detected as an "application/msword" (Table 2).

Table 2. Eight possible matches for a File with a ".doc"-File Extension

PUID	Format	Version
x-fmt/2	MS Word for Macintosh	6.0
x-fmt/42	Wordperfect Secondary File	5.0
x-fmt/43	Wordperfect Secondary File	5.1/5.2
x-fmt/129	MS Word for Macintosh	X
x-fmt/131	Stationery for MAC OS X	
x-fmt/273	MS Word for MS-DOS	3.0
x-fmt/329	Interleaf Document	
fmt/609	Microsoft Word (Generic)	6.0-2003

Much worse in this case, the file was not accessible anymore and opening it with a Hex viewer did not give a hint about the actual file format, either. Of course it could be a binary file the staff from the Indiana University Library [6] had to deal with as well, which was a false positive for MS Word because of the ".doc"-file extension.

Group 6: Executables
Examples: DOS-exe, Win-exe, Java-class
Archived Format: original
Known Issues: still more research needed

There are other formats not included in these six groups, e. g. Audio Formats (.cda), Video Formats (.mpg, .flv) and many more, which need more attention and research until a preservation decision can be made. All in all, there are 177 different file formats in our sample, 120 of them occur less than 10 times, which makes it difficult to examine the formats in detail. We aim to add these formats to the Pronom Library, but it would be helpful if the sample would be bigger.

Lessons learned

Some of our data carriers contain more than 2,000 files and the file formats are heterogeneous, some are not part of the Pronom Format Library used by DROID yet and some file formats that are generally known to Pronom cannot be detected because of issues with the file quality. Even if only one file of the archive package has issues, the ingest process of the whole package stops and needs manual assessment. This is one of the reason why the NLNZ staff spend some time during pre-ingest for preconditioning the files, solving most of the problems before the actual ingest starts [4].

One of our CD ROMs that contains 1,021 files – which is not unusual – managed to create three different error alerts for 64 different files. Two did not make it through the virus check, for four of them the technical metadata could not be extracted and 55 others failed the JHOVE format validation.

As a workaround to ensure at least bit stream preservation for the CD ROMs with "difficult" files, we have decided to allow unknown formats in our archive as well, as long as the items are well described and easily searchable for later improvement actions.

Conclusion and Future Work

The ingest workflow is easily embedded in the daily work of the Digital Archivist. As the workflow is documented in detail, it can be delegated to other staff members. The information about file formats and number of files is automatically extracted from the original data carrier and so is the information about the original structure. However, as only parts of the workflow are automated, the scalability has its limits, especially because the hands-on work on the archival packages with errors has proven to be time-consuming.

Besides, it might be a good idea to follow the example of the NLNZ [4] to solve issues with the files and file formats during the pre-ingest stage. Nevertheless, it will be necessary to document all edit actions, changes and information extractions in the metadata within the Rosetta archive.

Experience has shown that JHOVE considers most "unusual" files as a "well-formed and valid bytestream", which is not as specific as we would wish. The workflow has shown some blind spots for JHOVE and DROID. We want to share our findings and experience with the community. Maybe we could contribute to international efforts of tool building and add to already existing format libraries like Pronom. The publishing of experiences and workflows in progress in blogs [3] [5] has been very useful for us during the development of the workflow, as has the communication via twitter and email to share experience and information among colleagues.

The test of curation workflows like migration of obsolete formats is next on our list.

We do not provide access to the archived material from the external data carriers yet, because our users still access the data directly from the CD ROMs in the stacks. This is due to change once the copies in our archive are needed because the access to the data carrier in the stack is no longer usable. Unfortunately, there are not only organizational and technical issues to be solved, but legal ones due to copyright restrictions as well.

The golden – and most user-friendly – way would be a web-based access to the content of the CD ROM, like the one the Indiana University Library provides to their users [6]. This is only possible if the copyright provisions of the material allow free access via web, which is the case for the material selected by the Indiana University [6]. For our material the situation is different, the material is copyright-restricted and in most cases will remain so for the next 70-plus years.

References

[1] A. Dappert et al, "Developing a robust migrations workflow for Preserving and curating hand-held media," iPRES (2011).
[2] D. Elford et al, "Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia" IFLA (2008).
[3] J. Mitchum, "A short detective story involving 5,24 inch floppy disks," Digital Archiving at the University of York, Blogpost (2013).
[4] L. Rosin, "Applying theoretical archival principles and policies to actual born digital collections," National Library of New Zealand (2013).
[5] T. Sant. "Establishing a Workflow Model for Audio CD Preservation," OPF Blog (2013).
[6] K. Woods et al, "Creating Virtual CD-ROM Collections," The International Journal of Digital Curation (2009).

Author Biography

Having graduated from the Humboldt University Berlin in 2011, as a Master of Library and Information Science, Yvonne Friese joined the Leibniz Information Centre for Economics in Kiel as a Digital Archivist. She is an active member of nestor, the German network for Digital Preservation and leads two nestor working groups: preservation policies and costs of Digital Preservation.