Survey and assessment of sources of information on file formats and software documentation

Final Report

The Representation and Rendering Project

University of Leeds

The Representation and Rendering Project is funded by the Joint Information Systems Committee (UK)

Contents

1.0 Executive Summary

Advancing technology is threatening the survival of digital materials. As hardware and software become obsolete, the ability to view or render a digital object can be lost. The effective lifetime of digital objects can be as short as 5 or 10 years. A wide range of solutions are being developed to enable digital objects to be rendered or viewed when the original hardware and software with which those objects were created is no longer available to the user. Without these rendering solutions the many file formats used to encode digital objects will be meaningless to future users.

Rendering solutions involve the migration of data from the original file format to a different format or the emulation of the hardware or software that renders the data. The former requires knowledge of the file format in question and the latter requires information about the hardware, operating system or application software to be emulated. Documentation in these key areas is not always made open by the companies which develop the file formats, software or hardware.

Collecting file format and software documentation from the wide variety of sources available is an essential step in enabling the development of digital preservation solutions. This documentation must be preserved over time and combined with knowledge of available rendering solutions to provide effective long term digital preservation.

2.0 Introduction

2.1 Background

Software and hardware obsolescence is increasingly regarded as a dangerous threat to the survival of digital materials.

Initiatives like the UK's Digital Preservation Coalition [1] have taken the lead in highlighting the risks facing our digital heritage. Examples like the BBC Domesday Project [2] have shown that action must be taken to preserve digital materials before it is too late.

2.11 The problem

At their most basic level, digital objects are sequences of zeros and ones which represent encoded data. Different file formats specify how these codes represent the intellectual content created by a digital object's author. An example of which is the Microsoft Word format. This format is a specification for the storage of textual data, along with formatting information. Most file formats are incredibly complex, making the codes meaningless to a human observer. In order to make sense of a digital object, software is required to interpret and display or render the data for the user. In the example of the Microsoft Word format, Wordview could be used to render this file format.

Application software is used to create and edit most digital objects. Around the time of creation of a digital object the same application software can be used to render and view that object. Unfortunately, a range of factors can lead to the loss of this rendering ability. Rapidly advancing technology and the obsolescence of computer platforms or operating systems leaves users without the ability to run their older application software. New versions of application software may not support earlier file format versions. Software developers may go out of business and no longer support the applications they developed. To the casual user this may not seem like a major threat, but in the space of just a few years it can become impractical, difficult or even impossible to render a digital object.

2.12 Migration using application software

Most application software provides a degree of backward compatibility, enabling older file format versions to be imported into current application software. In some cases this can be used to migrate digital objects in danger of obsolescence to more current file formats. This has been suggested as a solution for preserving repositories of digital objects but a range of problems hinder the practicality of this technique.

Versions of file formats tend to be short lived due to the commercial interests of the software developers. Application software does not generally provide import facilities for every previous file format version. Only a number of recent formats will be supported. This forces migration from format to format to be performed at frequent

intervals. For a small number of digital objects this is certainly possible, but for a large repository of objects this can be costly and impractical.

Relying on what are usually commercial software developers to provide future migration paths can be a dangerous policy. If a developer goes out of business, finding a way to continue the migration path could become difficult, costly or even impossible.

Regardless of these problems there has always been a degree of doubt as to the effectiveness of this strategy in terms of maintaining the accuracy of digital objects when they are migrated between file format versions. Preliminary results from the Testbed Digitale Bewaring project [3] suggest that migration of this kind could lead to a considerable loss of accuracy. The project has been testing the migration capabilities of a range of available software, using painstaking visual comparisons between the original and the end results. Very simple objects could generally be migrated successfully, but more complex elements often failed to survive the migration process. Migration can therefore be a successful strategy for simple objects but is problematic for the more complex objects that are a significant percentage of products from modern software applications.

If migration of this kind leads to frequent loss of information, this whole strategy must be called into question for more complex objects. Performing migration with an application produces no record log of which elements have been successfully migrated and which have been lost or incorrectly migrated. With no record of this kind, the longevity of the objects to be migrated is very uncertain. The potential of significant loss over just a few migration iterations is great.

2.13 Preservation strategies

Alternatives to application software migration are being developed by a range of projects and institutions in the preservation field. Some involve the preservation of the original bytestream of a digital object along with specially developed rendering tools to migrate the object on request or render it using a viewer. Migration to a standard or open format like XML provides a different approach that still relies on the basic principles of migration. If a rendering in the original environment of creation is required, an emulator could provide the means to run the original application software on a current computer platform. The “CAMiLEON guide to new digital preservation strategies” [4] and the “Migration: Context and Current Status” [5] reports provide a more detailed description of some of these strategies.

2.14 Code books

It seems likely that a range of preservation strategies will be required to ensure the long term preservation of digital objects. Regardless of which of these alternatives is used to perform preservation, the lack of documentation may prevent successful action from being taken. Most application software developers produce file format documentation for the formats they design and develop. Not all of them make this documentation available and even if they do, it is not always accurate. As has been

described above, digital objects are coded sequences. Without a code book to explain the meaning of these sequences, the vital rendering tools that enable the use of the objects over time cannot be developed. Reverse engineering of software or the digital objects themselves can provide some answers, although legal constraints may well prevent this kind of action. Even where reverse engineering is possible, without any file format documentation, the process is likely to be too laborious and expensive.

2.15 An example of the problem : Illustrator and Freehand

The Los Angles Times maintains a repository of the digital objects from which its newspaper is constructed. Rapidly advancing computer hardware and computer software is causing major problems in the access to and re-use of these digital objects. Objects held in the repository include many thousands of line drawings in Adobe Illustrator, Macromedia Freehand and Macdraw formats. The two main aims of the repository and the difficulties associated with achieving them are illustrated with a topical example from the 1991 Gulf War.

The newspaper wishes to maintain a historical electronic record and the ability to re-use material in the newspaper where required. 12 years on from the first Gulf War, the current war with Iraq highlights the importance of both of these requirements. The following examples are screenshots of real images from the paper.

Victoria McCargar Senior Editor / Library Projects at the LA Times, describes the first example *"The graphic was created in FreeHand 1, migrated to FreeHand 5 and then to FreeHand 9. The original font calls were broken, and the default replacement typeface differed in letter spacing, word spacing and kerning pairs. The vector geometry--the drawing portion--migrated intact."* It should be noted that the current version of Freehand will only import versions 7 and above. This leaves a complex multiple application migration from version 1 to the latest version on request. Alternatively migration can be performed from version to version of all objects in the repository every time a new version of Freehand is released. A single remaining copy of Freehand version 5 was traced and put into action to get access to the version 1 objects created at the time of the first Gulf War. Many problems were encountered with fonts, missing photographs and drawing specifics.

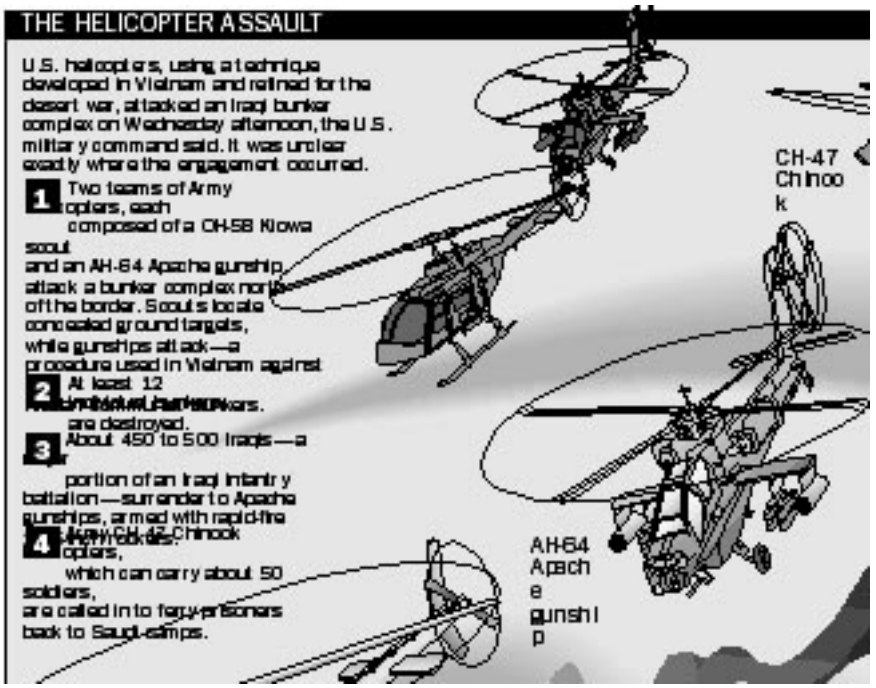


Figure 1 : Screenshot of Freehand drawing, following lengthy path of migrations

Figure 2 shows a screenshot of a graphic migrated using a lengthy path from its original state as an EPS file saved from an obsolete version of Illustrator. This is viewed in a recent version of Freehand. Several problems have been introduced by the migration steps. The font and some of the line elements have lost accuracy and become pixelated, and the green fill was originally a dithered green which appears to have been scaled up.

Figure 2 : Screenshot of Illustrator EPS following migration to Freehand

Figure 3 shows the difficulties of rendering when application based migration paths simply aren't available. A graphic originally created in MacDraw became obsolete and wouldn't open in other applications like Illustrator. As McCargar describes *"We were able to locate a copy of MacDraw and install it on my Macintosh G3. The graphic opened, but it rendered thus."*

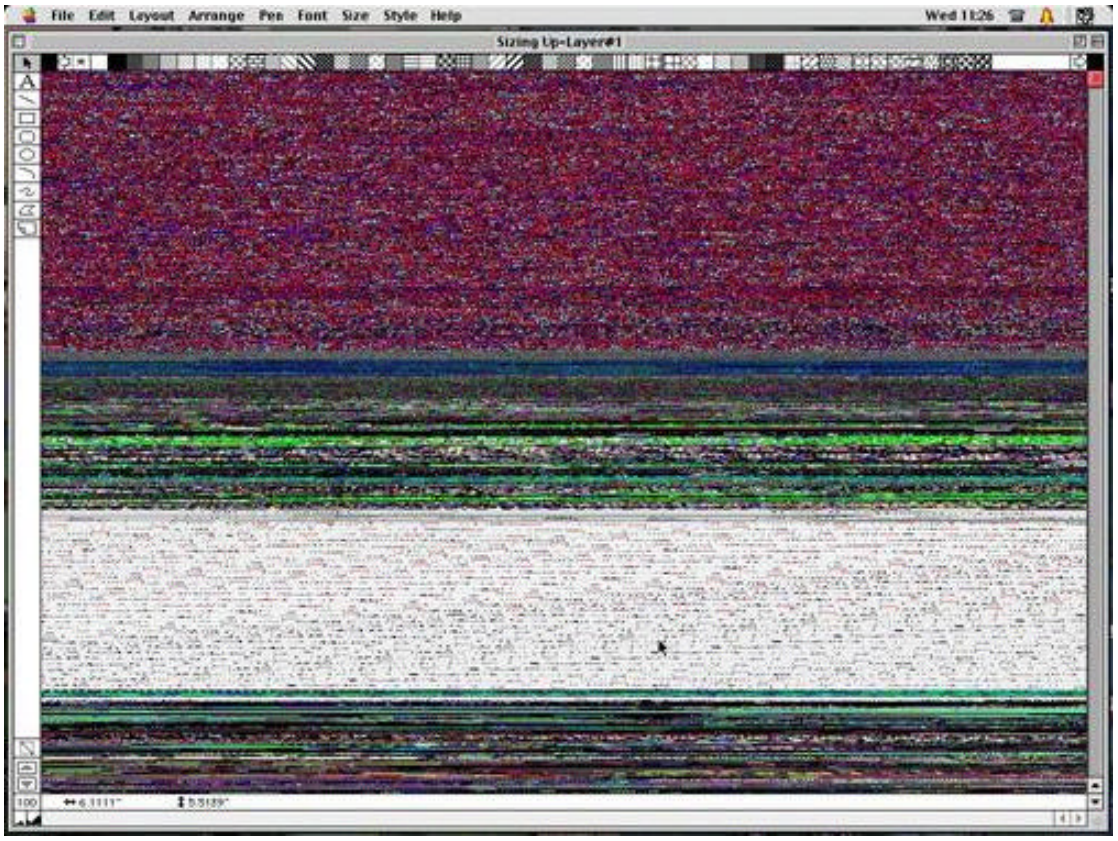


Figure 3 : Attempting to view a graphic in MacDraw

A correct rendering in the original MacDraw software was possible if the screen resolution of the modern Mac it was running on was reduced to 640x480! Given the current rendering capabilities at the LA Times archive, this object is teetering on the brink of obsolescence. Only new migration facilities or an emulation solution could prevent it from being lost.

The last example shows a recent object placed into the archive (see figure 4). McCargar describes *"It's interesting, because the mountains are created as a Photoshop EPS and embedded in the vector file. We deal with these compound objects all the time. It's not unusual to have a vector graphic that in its native, i.e., reusable, format comprises as many as five or six other images as well as auxiliary files such as spreadsheets or GIS map data. At the moment we are archiving all of it."* This highlights the increasing range and complexity of digital objects which will require preservation solutions in the very near future.

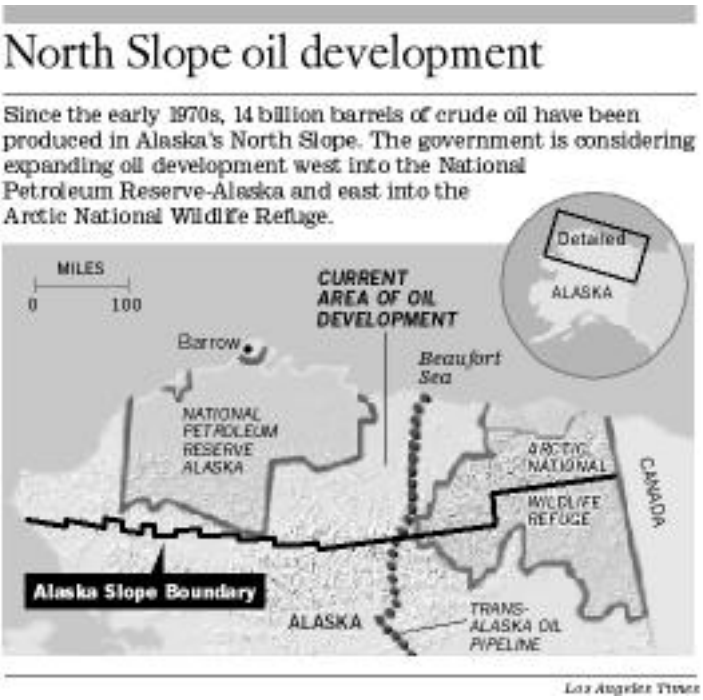


Figure 4 : Screenshot of typical compound vector graphic

These examples illustrate very clearly the potential that digital preservation can play in maintaining a historical record and facilitating commercial advantage through re-use. The design and implementation of rendering tools to enable this can only be put into practice if sufficient file format documentation is available to decode aging and obsolete formats. The quality and availability of documentation on the file formats discussed above varies tremendously, and will be described in detail below.

3.0 Aims and objectives

The survey will focus on 5 key aims:

1. Discovery of sources of file format information
2. Analysis of the extent of available file format information
3. Analysis of the accuracy of available file format information
4. Suitability of available file format information for the development of preservation tools
5. Overall analysis of the impact of the availability (or otherwise) of file format information on the tasks facing the digital preservation community.

The overall objective of the survey is to assess how the availability of file format information may affect the development of tools and strategies to enable digital preservation to be successfully performed. Recommendations will be made to JISC in order to inform its strategy for digital preservation and the forthcoming "Digital Curation Centre".

4.0 Scope

The survey has a short timescale for completion and hence will adhere tightly to the aims and objectives listed above. The survey will not aim to gather and record extensive amounts of file format information. This is seen as the role of prospective file format registries or Representation Information systems (see below).

This short survey should not be considered exhaustive, indeed many sources have not been recorded. The survey will instead aim to provide a representative sample of the range of sources available. Where appropriate, examples will focus on text based digital objects which is the same field to be addressed by current Migration on Request work on the Representation and Rendering Project. Text based objects are a crucial part of the spectrum of digital resources to which preservation solutions will need to found. They therefore provide a useful specific example of the problems being faced across the board. By illustrating the discussions with text based examples throughout the report, comparisons can be made between the differing sources of file format information.

5.0 File format information

The following categories for information have been assessed:

- Web site collections
- Funded collections and sources
- Research and development
- Open source developers and software
- Commercial developers and software
- Published sources
- File format identification
- Software documentation

5.1 Web site collections

The open source and enthusiast community has for some time recognized the need to gather and provide file format documentation. A number of enthusiasts maintain web sites devoted to collecting and giving access to file format specifications.

Wotzit's Format? [6]
http://www.wotsit.org/default.asp

Wotzit is generally regarded as one of the best file format web sites and the first place to begin a search for specific file format information. A large number of file formats are covered and a search facility is provided by the site. The maintainer of the site was contacted but did not respond. There is no evidence of how recently the site has been updated other than the copyright notice which reads *"This site is © Paul Oliver 1996-2002."* [6]

MyFileFormats [7]
http://myfileformats.com/

Myfileformats is another very useful source of documentation, but again no response was received from the maintainer and the only evidence of current updates was the copyright notice which reads *"© 2000-02 MyFileFormats.com."* [7] Some problems were encountered with the site which did not always work reliably over the period that this survey was conducted. Much of the functionality listed by the site on the homepage is not currently provided.

File Format Encyclopaedia [8]
http://pipin.tmd.ns.ac.yu/extra/fileformat/

The File Format Encyclopaedia is another collection of file format documentation. Limited contact was made with the maintainer of the site.

Sample from FFE:

Text

ANS	Ansi Escape Sequences
DOC	Pilot standard text document file format
	Wordperfect File Format by Max Maischein
	File Format for WordStar Release 7.0
	Microsoft Word 6.0 Binary File Format
	Microsoft Word 97 Binary File Format
	Microsoft Word 97 Binary File Format NEW
	XDOC Data Format
EPS	Adobe EPS 1.2
	Adobe EPS 2.0
	Adobe EPS 3.0
HTML	HTML 3.2 Reference Specification
PDF	PDF Reference second edition Version 1.3, Adobe Systems Incorporated
RTF	Rich Text Format (RTF) Specification, version 1.6
	Rich Text Format (RTF) Specification
	Word Rich Text Format (RTF) Addendum
WRI	Write File Format

The documentation on the FFE site is all held locally on that site, rather than simply listing links to other sources. The information varies greatly from item to item. The Pilot document contains minimal information on the format as does the Wordperfect document which only describes Wordperfect related file types and their common headers. This information is useful for automatically identifying file types of these kinds but not for actually rendering them. The Word 6.0 document is large and detailed and presumably sourced from Microsoft itself. The top of the document is labelled *"Microsoft Confidential"* and dated 09/03/94.

Szuper [9]
http://www.szuper.biz/

The Szuper site collects a range of computer related information, including an unstructured list of file format documents. The site lists 173 different formats with accompanying documentation. As with the other web site collections described above, the source and accuracy of most of the information is unclear. Some of the documentation is written by enthusiasts who have begun to reverse engineer and document some aspects of a number of file formats. Much of the same information can be found on the other web sites described above.

Sample from file format list on Szuper:

> …
> *34 DOC Méret : 155K Nyelv : Eng Forráskód : -*
> *The WordPerfect file format for WordPerfect 5.0/5.1 41*

35 DOC Méret : 2K Nyelv : Eng Forráskód : -
Microsoft Word for Windows 6.0 Binary File Format. Updated struct 58

36 DOC Méret : 86K Nyelv : Eng Forráskód : -
Microsoft Word 6.0 Binary File Format 56

37 DOC Méret : 77K Nyelv : Eng Forráskód : -
Microsoft Word for Windows 6.0 Binary File Format 09/03/94 53

38 DOC Méret : 89K Nyelv : Eng Forráskód : -
Microsoft Word 97 (aka Version 8) Binary File Format. Revised Aug 1 1998 58

39 DWARF Méret : 118K Nyelv : Eng Forráskód : -
DWARF Debugging Information Format. Industry Review Draft. Copyright ó 1992, 1993 UNIX International, Inc. 34

40 DWG Méret : 2K Nyelv : Eng Forráskód : -
AutoCAD DWG (R12) file format 72

41 DXF Méret : 24K Nyelv : Eng Forráskód : -
AutoCAD's ASCII Drawing Interchange (.DXF) Files. Binary Drawing Interchange (.DXB) Files 62

42 ELF Méret : 74K Nyelv : Eng Forráskód : -
Executable and Linkable Format Specification, V1.1 46

43 EMD Méret : 4K Nyelv : Eng Forráskód : -
EMD module/song format for Advanced 16-Bit Tracker (ABT) 38
…

Further sites

There are many more web sites listing similar file format information. A number of further examples are shown below:

- WhatIs? http://whatis.techtarget.com/
- Almost every file format in the world! http://www.ace.net.nz/tech/TechFileFormat.html
- File extensions http://www.icdatamaster.com/

Analysis : Web site collections

Web site collections are likely to be valuable sources of file format information for use in the digital preservation field. Information is provided on many file formats but coverage is patchy and original sources usually unclear.

The lack of permanence of the sites is of particular cause for concern and highlights the need for the collection and maintenance of this information by the preservation community. A number of web based FAQs list many file format collections on the web, but the majority of these are broken links. It is clear that the future existence of these file format collections cannot be relied upon.

The legal ownership of the documentation on these sites is in most cases unclear. Some of the documents appeared to be originally intended as internal to commercial developers. This may cause problems for file format registries which wish to gather, preserve and make available documentation of this kind.

5.2 Funded collections and sources

The Diffuse Project and OII [10]
http:///ww.diffuse.org

The project web site describes the aim of Diffuse as *"…to provide a single, value-added, entry point to up-to-date reference and guidance information on available and emerging standards and specifications that facilitate the electronic exchange of information."* [10] Some file format information and classification can be found although it does not go into any great detail. Related information which lists and describes standards bodies can also be found. There is little analysis or opinion on the various formats and standards that are described, but Diffuse provides some pointers for the information which will need to be recorded in digital preservation file format registries.

As stated on the homepage, the permanence of the site cannot be relied upon. *"The Diffuse project, which built on one of the first online services launched by the European Commission in 1995, concluded on 31st January 2003. No decision regarding maintenance of the contents on this website has yet been made."* [10]

File format information can also be found in the Open Information Interchange initiative [11] (part of Diffuse) which includes specific information on file formats suitable for interchange and sharing.

http://www.diffuse.org/oii/en/oiistand.html#oiistand

Sample from OII:

> ***TeX DVI***
>
> ***Expanded name***
> *TeX Device Independent File Format*
>
> ***Area covered***
> *Language used to interchange files formatted using the TeX formatting language between different output devices.*

Sponsoring body and standard details
Proprietary specification, developed by Donald Knuth of Stanford University, which is distributed under the auspices of the American Mathematical Society

Characteristics/description
Output produced by TeX formatters for interchange between printing devices. Commonly used to produce mathematical and other scientific texts.

The TeX primitives provide a very powerful set of typesetting controls. They are, however, difficult to use in their raw form as they form a fully-fledged programming language, which includes facilities for defining your own character shapes. As TeX also provides powerful facilities for compiling structured sets of macros most users generate documents that are coded using TeX macro sets, of which LaTeX is by far the most popular.

Work is currently underway to extend LaTeX to provide the type of facilities typically provided through SGML and HyTime. The SIMSIM TeX macro package can be used to convert SGML documents into TeX format.

Usage (Market segment and penetration)
TeX is widely used in the academic environment. For document interchange most people exchange unformatted TeX files, but this can lead to problems if different macro packages (or different versions of the same package) have been used. DVI files avoid this problem as the document is interchanged in its formatted form, though it should be noted that the DVI file contains no information about the fonts to be used to reproduce the file, relying on the receiver having the same fonts as the document's generator. Another problem is that the DVI file can be considerably longer than the source document!

Further details available from:
American Mathematical Society, 201 Charles Street, Providence, Rhode Island 02904, USA

AMS maintain a WWW page that points to a wide range of TeX resources at http://www.ams.org/tex.

National Software Reference Library [12]
http://www.nsrl.nist.gov/index.html

The National Software Reference Library (NSRL) is described on the web site as aiming to *"…collect software from various sources and incorporate file profiles computed from this software into a Reference Data Set (RDS) of information. The RDS can be used by law enforcement, government, and industry organizations to review files on a computer by matching file profiles in the RDS. This will help alleviate much of the effort involved in determining which files are important as evidence on computers or file systems that have been seized as part of criminal investigations."* [12]

This source could equally be of use to the digital preservation community and it seems possible that the National Institute of Standards and Technology (NIST) would be interested in collaboration with file format registry initiatives. The level of detail

describing each file format is somewhat limited and falls far short of any real file format specifications, but the RDS may be useful in the identification and classification of data formats and rendering software.

The NSRL Reference Data Set can be purchased for 90$ per year and a specification of the RDS itself can be downloaded from the NSRL web site.

PDF Zone [13]
http://www.pdfzone.com/

PDF Zone is a detailed resource of documentation, hints tips, tools and discussion focusing on the PDF format. The web site describes its mission as being *"…dedicated to energizing the PDF community with unbiased news reporting, timely interviews, expert resources and vibrant discussions."* [13] The independence of the site is unclear but the range of information makes it a valuable resource.

5.3 Research and Development

A wide range of organisations have conducted research into digital preservation issues related to file format documentation. A sample of this work is described here. Few researchers have expertise in a range of file formats, but many have expertise in one or two. Drawing on this existing expertise, either in the academic community or the commercial world where specific file format problems may have been encountered, will be essential for the future development of rendering tools.

Archiving and Preserving PDF Files, John Mark Ockerbloom [14]
http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2

Ockerbloom provides a detailed study of the issues surrounding the preservation of the PDF format. In particular, he discusses the future development of the format and when and where things might go wrong. *"Adobe is highly invested in the success of PDF, but even if Adobe fails or abruptly changes course, a community of third-party tools for handling PDF has started to emerge. PDF is used widely by many well-funded bodies (including the U.S. government, which is now using PDF as its standard way of distributing government publications) so there should be widespread support for using and migrating PDF, should Adobe fail to provide adequate support for the format.*

Even so, it is likely that PDF will one day be superseded by another format. It may be a successor format (as PDF is to Postscript), or it may be a completely different format that users prefer over PDF. Hence, it is necessary to have migration strategies planned for PDF." [14] The later discussion focuses on the limited ways in which information can be extracted from a PDF file. Currently there is no ideal destination format to migrate to from PDF and this increasingly leaves users more dependent on Adobe and PDF support. *"RTF (Rich Text Format) may be the best target for word-processor-oriented conversions at this time, though it is far from a perfect choice."* [14]

Cartagis Survey [15]
http://www.cartagis.com/fileformatsurvey.html

Cartagis recently conducted a survey of users in advance of the development of a new application to perform Spatial Analysis on the Apple Mac. This is a very specific use case, but provides some very interesting insights into software users and the drivers behind their choices of data creation and use. The survey results shown on the site include graphs of the preferred input and output formats to a new GIS application. The results show a wide ranging spread for both the import and export of data. The overall preference was for the ESRI Shapefile format with Arcinfo not far behind. These formats are leaders in their field and are reasonably open with good support from open source developers (see [16]). The popularity of other formats is also significant. The users suggested a preference for many different options for import and export, many of which are proprietary and complex. In such a niche area of data representation, this reveals a lot about the motives of users in the choice of format that their documents are created in. Wheatley discusses the influence of commercial drivers on document creation in "The CAMiLEON Guide to New Digital Preservation Strategies" [17].

Data Curation for e-science in the UK (JISC)
http://www.jisc.ac.uk/index.cfm?name=project_escience

The E Science curation audit draft report from JISC provides some interesting responses from the audit questionnaire to researchers.

> *"Question 8: Will future users need any of the following to use the data?*
> *Yes No*
> *a) Special software*
> *20 (42%)*
> *28 (58%)*
>
> *b) Special hardware/instrumentation*
> *7 (15%)*
> *41 (85%)*
>
> *c) Explanatory documentation*
> *36 (75%)*
> *12 (25%)*
>
> *A disturbingly high number of respondent's report using special software whose longevity may be questionable.*
>
> *Some of the notes made are as follows:*
> *'We use European Data Format EDF for patient data which is standard.'*
> *'The special software is not essential but it is highly desirable.'*
> *'Specific yes, Special no.'*

'Some data yes, most no.'
'Software: probably, but unlikely to be of much use dependent on time since data produced. H/w/instrumentation: Probably but unlikely to be available again dependent on time since data produced. Other: a mechanism to restore archived data and re-archive in the new format to maintain its usefulness.'
'Re special software: some aspects of the processed data would be far easier to access with specialised software, but most data is in standard formats.'
'My belief is that the material archived should be self explanatory and accessible.'
'Explanatory documentation, because nearly all qualitative data needs to be placed in context.'
'End users will not require these at their own sites, but it will be necessary either to maintain the database server and database management software at the site where the data are held, or to arrange export of the data in a 'universal' format when the project comes to an end. Much of the value of the data would thereby be lost because it is currently maintained in an object-oriented database with sophisticated linkages.'

Question 10:
(a) What are the main types of specialist commercial or open source software you are using, if any (e.g. Maple, TurboChrome, etc)? :
This elicited a surprising variety of software systems - from this sample of 29 replies some 49 systems were mentioned, most of them just once. (This counts a reply such as 'Microsoft Office' as one system.) Of those mentioned more than once, MatLab was mentioned 7 times and Fortran (sic) and MS-Office both 4 times; a further nine were mentioned twice.
(b) Are you using software you have written for your project?
Yes 22 (46%)
No 26 (54%)

This recalls question 8a, whether special software is in use. It is not clear how this self-generated software is to be preserved if need be, and a nearly 50% response indicates that the potential problem may be of considerable size."
[68]

The draft report highlights the two key issues for preservation in an E Science context, that of size and of specialist software.

PDF-Archive [18]
http://www.aiim.org/standards.asp?ID=25013

PDF/A or PDF-Archive is an encouraging initiative to specify a subset of PDF tags for archival purposes as an ISO standard. This is currently in its second draft. Although PDF has seen widespread take up across the preservation and archival communities, there are problems associated with some of the more complex PDF tags as well as issues surrounding availability and use of fonts. The PDF-Archive standard will hopefully provide and agreed specification to which PDFs should be created. The provision of tools for verification and possibly migration from non PDF/A documents will be crucial.

Migration on Request – CAMiLEON [19]
http://www.si.umich.edu/CAMILEON/reports/mor/index.html

Migration on Request is a concept developed by the Cedars and CAMiLEON projects for the preservation of non-interactive digital objects. The associated papers provide some interesting discussion of file format documentation and the issues surrounding the selection of file formats for proof of concept testing of preservation strategies.

The Florida Center for Library Automation (FCLA) [20]
http://www.fcla.edu/digitalArchive/index.htm

The Florida Center for Library Automation (FCLA) has produced some invaluable investigations and analysis of file format preservation issues. The first 5 of the FCLA's action plans are available for download from the web site.

The web site describes some of the preservation actions the FCLA is taking *"The FCLA Digital Archive will accept submission packages from participating partners, ingest digital documents along with the appropriate metadata, and safely store on-site and off-site copies of the files. An action plan will be developed for each file format, which might include the creation of canonical derivative versions and/or format migration. A primary characteristic of the FCLA digital archive is its role in serving the real needs of a diverse group of libraries. Every effort will be made to accommodate the formats important to the institutions, whether these are traditionally considered "archival" or not."*

Andrea Goethals of the FCLA describes the sources for the actions plans. *"The action plan background reports are the background information culled from the specs and online papers/articles that can help decide how 'stable' a format is and what the distinguishing characteristics of a format are. The action plans describe the long- and short-term actions we'll take regarding a particular file format."*

These plans provide a detailed breakdown of the issues involved in preserving specific file formats. The PDF background plan provides some particularly interesting observations, including a bar chart of the size of the PDF specification in pages at each version, and the length of time that each PDF version has survived.

Risk Management of Digital Information:A File Format Investigation, Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger, William H. Walters, Anne R. Kenney, June 2000 [21]
http://www.clir.org/pubs/reports/pub93/contents.html

The File Format Investigation undertaken by CLIR provided a seminal analysis of the difficult issues surrounding migration between file formats and the associated risks of performing these preservation actions. The report is an invaluable description of the team's test case work with a number of formats including Lotus 123 and TIFF.

The Investigation revealed errors in existing file format documentation and recognised that these mistakes had been repeated by a number of other sources (see Printed Sources, below). This example reinforces the fact that even where documentation for a specific format can be obtained from a reliable source, there will always be omissions and mistakes.

The report notes that *"Since basic file structure concepts are common to many file formats, experience with one format can be used to understand other formats."* [21] While it is accepted that documentation will still be required in order to understand the "other formats" in question, the similarity of data structures can be invaluable in their interpretation. The concept of Migration on Request [19] (described above) exploits this fact by providing one rendering tool for a number of file formats of similar type. From a broader perspective, taking advantage of skills and experiences in specific file formats will be an important challenge for the preservation community.

The Investigation showed that risks involved in migration can be identified and measured. As the report states *"The greatest challenge is the interpretation of the risk, i.e., to determine when a risk is acceptable. Risk-assessment tools cannot replace experience and good judgment. The tools can be compared with navigation aids used on the high seas. Following five centuries of intensive effort to develop risk-reducing technologies, ships' helms are still manned, and collisions between ships at sea still occur."* [21] The same analysis holds true from a broader perspective of identifying the risks associated with the dependence of data on specific formats and the level of documentation associated with those formats.

"Public Draft -Extracts from A Survey of Information Technology Vendors",
Philip Lord [22]
http://www.dpconline.org/graphics/reports/

A recent survey of IT vendors conducted by Philip Lord for the Digital Preservation Coalition touched on the subject of access to file format information. Lord asked respondents whether they felt they could contribute to a register of file formats. Most of the responses were positive, but not in the actual contribution of information about their own file formats. As Lord describes *"Perhaps significantly, no respondents offered to submit their own file formats to such a register; most offers were on peripheral services; or they claimed, probably with justification, that they were not in control of this information. Companies who are producing software were reluctant to contribute, as commercial ownership and competitive advantage are obstacles. However one company suggested that contributing to a register with a five-year 'moving wall' would be acceptable, where formats over five years old were contributed; five years is a long time in the software business and such formats would be effectively out of date."* [22] This last point is crucial. It is vital to perform preservation work before it is too late. A compromise between obtaining the required file format information when developers are happy to release it and implementing the required preservation tools before it is too late may well be possible.

Lord completes this section of the report with an ominous quote from a hardware vendor *"The real formats are the property of the monopoly supplier. They have*

strong commercial reasons to keep these formats to themselves, and it would be a doomed effort. Government mandating of open standards would help but would be followed by massive lobbying." [22]

xanadu archive - Xml Application for Normalising, Archiving and Displaying Universally
http://sourceforge.net/projects/xanadu-archive/

Xanadu is a tool developed by the National Archives of Australia to convert electronic records into XML, and to view the converted XML documents. It is a Java 2 based GUI application, and is open source; the first release version is expected mid 2003. The NAA created xanadu for their own use and plan to use it for all their XML migration work, and hope other agencies will use it for migrating towards open XML based formats.

"We've designed xanadu so it has a plug-in architecture. This means that the main xanadu application handles the user interface, but all the converting into XML and viewing XML is handled by little pieces of Java code especially designed for particular document formats.

xanadu is able to convert documents from a 'source' format into a 'preservation' format only after a special plug-in has been written containing the conversion rules for that particular migration. So xanadu can convert comma separated text into an XML dataset document, or can convert the recordset returned by a SQL query into an XML dataset document because we have written a plug-in to handle these two conversions. But at the moment, xanadu cannot convert an AmiPro document into an OpenOffice document because we have not yet written that particular conversion plug-in.

One of the big advantages of the plug-in architecture is that anyone who can write Java code can write small plug-in applications that extend the number of document formats xanadu can convert or can view." [66]

5.4 Open Source Developers and Software

The popularity of Open Source software is rapidly increasing and this enthusiast driven market offers much of use to the preservation community. Open Source software can provide useful rendering solutions as well as valuable sources of file format information. These sources can include documentation produced by developers and made available to others, human experience from the developers themselves who have a working knowledge of specific formats, and open source code which processes digital objects.

PDFbox [23]
http://www.pdfbox.org

PDFBox is a Java Library providing functionality for manipulating PDF documents. As the web site describes *"This project will allow access to all of the components in a PDF document. More PDF manipulation features will be added as the project matures. This ships with a utility to take a PDF document and output a text file."* [23]

Currently the library only provides a very basic text extraction but shows promise for more complex development. The open source libraries in addition to the detailed documentation and example code provide a very useful source of information on the PDF format.

Xpdf [24]
http://www.foolabs.com/xpdf/

Xpdf is the best of the Open Source PDF viewers available for free use. A range of related utilities are also provided, including a PDF text extractor and a PDF-to-PostScript converter. As the web site describes *"Xpdf runs under the X Window System on UNIX, VMS, and OS/2. The non-X components (pdftops, pdftotext, etc.) also run on Win32 systems and should run on pretty much any system with a decent C++ compiler."* [24]

Xpdf highlights the success of Adobe's open strategy towards the development and uptake of the PDF format, as one of the best PDF viewers available.

Author of Xpdf, Derek Noonburg, provides some valuable insights into PDF, Adobe's strategies and the accuracy of their documentation.

"Adobe's PDF spec has gotten better since the original release (PDF 1.0) the current version is 1.4). I've had to do some reverse engineering, mostly regarding corner cases, not completely undocumented sections. For example, there is an ambiguity in the case where a path consists of a single point -- what do stroke, fill, and clip do in this case?. I ended up constructing PDF files and checking to see what Acrobat Reader does."

"Handling of TrueType fonts is another case. The PDF spec says:

| Note: Some popular TrueType font programs contain incorrect encoding
| information. Implementations of TrueType font interpreters have
| evolved heuristics for dealing with such problems; those heuristics
| are not described here. For maximum portability, only well-formed
| TrueType font programs should be used in PDF files.
[PDF 1.4 spec, page 353]

Of course, there are lots of PDF files containing badly-formed TrueType fonts. Making Xpdf behave the same as Acrobat with these has been a hassle."

"When Adobe added encryption (for the security settings), they originally didn't release the spec. The problem is that it's a pure security-through-obscurity system -- once the viewer can decrypt a file, there's nothing forcing it to honor the no-copy/no-

print/etc. bits, other than Adobe's request to do so. They did eventually release the spec, even sending a draft to various open source developers, including myself. (The latest version, which added 128-bit encryption, had some ambiguities and errors, but Adobe has been making an effort to clean this up.)"

Noonburg was asked if he had been hampered in the development of Xpdf with any legal concerns *"The Unisys LZW patent was the only big one. Adobe has several patents, most of which are freely licensed to anyone writing PDF generator/consumer software (as long as you stick to the PDF spec). There is one patent relating to linearized 'optimized') PDF files which is licensed only for generators, not for consumers (viewers)."*

Adobe Legal Notices [25] describes the various patents held by Adobe and the terms and conditions associated with their use. Noonburg stated that there was nothing in particular that concerned him in this area other than the obvious possibility of Adobe reversing their policy and preventing third party support of PDF at some point in the future. This seems unlikely, but is always a possibility.

Noonburg was asked about Adobe's motivation for supporting open source developers. *"Adobe's business model with both PostScript and PDF includes support for third-party developers. They come up with a spec, and then make it really complex so that they (Adobe) have what is clearly the best implementation. If you look at PostScript printers, for example, there are at least a couple of non-Adobe PS interpreters being shipped, but they tend to be in lower price printers, and are generally perceived as being not quite as good as Real Adobe PostScript. In both PDF and PostScript, there are lots of niches that are too small for Adobe to bother with, and so they're happy to see other developers taking them. The more people using PDF, the more money Adobe will make from their own products.*

Regarding my 'really complex' comment, I should add that Adobe's documentation is among the best I've seen. It has some problems, but they generally fix them. And even with the problems, their specs blow away most of the other technical documentation I have to deal with.

My guess is that Adobe sees Xpdf as useful evidence that the PDF spec really is open and implementable by third parties. I provide (with help from the open source community) a PDF viewer for a lot of obscure platforms that Adobe can't possibly support, given cost tradeoffs of a commercial product."

Noonburg was asked if the changes made to the PDF specification had really been backward compatible or were fixes to Xpdf required as the PDF format was developed. "No, they add new features to the spec with each new version. They don't go back and break existing features. The latest versions of both Xpdf and Acrobat will happily read PDF 1.0 files. Xpdf doesn't have any special case code based on the version number in the file."

The take up of Xpdf provides some evidence of the accuracy and quality of the tool and it has been ported to a number of different platforms, also pointing to the potential for re-use of the source code. Author of the RISC OS version of PDF [26], Colin

Granville, was asked if Xpdf code could be re-used in the implementation of a PDF or PDF/A verification tool. *"It could but I think you'll find that you would strip out a lot of the code. After all you don't want to actually render anything and you are only interested in certain tags/commands. It obviously made it a lot easier for me to implement a pdf reader - I just needed the code that rendered to an output device."*

Open Office [27]
http://www.openoffice.org/

The OpenOffice web site describes an aim *"To create, as a community, the leading international office suite that will run on all major platforms and provide access to all functionality and data through open-component based APIs and an XML-based file format."* [27] This initiative is widely recognised as a very positive development that is particularly favourable to the digital preservation community. OpenOffice offers alternative and well supported software to the market leading Microsoft Office products, with the added confidence that comes with Open Source software. OpenOffice also provides an incredibly useful point of reference for developers wishing to support Microsoft formats. Useful information can be found in both the OpenOffice documentation and the source code itself. Almost all Open Source developers asked about OpenOffice thought that it was an invaluable source of file format information.

Wvware [28]
http://www.wvware.com/

Wvware is described on the web site as *"…a library which allows access to Microsoft Word files. It can load and parse Word 2000, 97, 95 and 6 file formats. (These are the file formats known internally as Word 9, 8, 7 and 6.) There is some support for reading earlier formats as well: Word 2 docs are converted to plaintext."* [28] Wvware provides the import facilities for the AbiWord open source word processor described below. The Wvware web site contains a comprehensive list of related resources, including tools for cracking password protected documents and a number of open source and commercial rendering tools.

Several small scale commercial companies offer services to decode password protected documents, of which CRAK [29] is one example.

LAOLA [30]
http://snake.cs.tu-berlin.de:8081/~schwartz/pmh/index.html

LAOLA is described as *"a collection of documentations and perl programs dealing with binary file formats of Windows program documents. LAOLA is giving access to the raw document streams of any program using "structured storage" technology to save its documents."* [30 ELSER is an offshoot of LAOLA which specifically addresses Word 6 and 7 formats. The software and source code comes with some useful technical guides and discussions.

AbiWord [31]
http://www.abisource.com/

AbiWord is an open source word processor, described on the web site as follows: *"Like most Open Source projects, AbiWord started as a cathedral, but has become more like a bazaar. AbiWord is part of a larger project known as AbiSource, which was started by the SourceGear Corporation. The goal of the project was the development of a cross-platform, Open Source office suite beginning with AbiWord, the project's word processor.*

SourceGear released the source code to AbiWord and a developer community quickly formed around the project. SourceGear has since then stopped work on the project, but still provides our servers and net connections, for which we are very grateful!"

AbiWord provides support for the import of various other file formats, including Microsoft Word.

Dom Lachowicz, a key developer of AbiWord and Wvware, was asked about the accuracy of file format documentation and how it affected the development of AbiWord. *"In practice, the documents listed on wvware and wotsit are largely complete and accurate. Sure, there are various undocumented, inaccurate, incomplete, and incoherent bits, but in practice this has turned out to be not so much of a problem. What has the potential to be a huge problem is that Microsoft no longer publishes such information on its formats - not even for reference on its MSDN site (which, btw, is an excellent and comprehensive resource), and that MSFT has since pulled all of its prior documentation on these formats.*

Older formats (such as Word2, Word5, etc) have limited support in wvware for various reasons, such as minimal or non-existent documentation on the format or that the format is grossly different than preceding or newer MSWord file formats (Word5 is nothing like Word97 or Word2. WordXP's format bears near-total similarity to Word97, though...). Also, very few of these documents are floating around cyberspace. Few, if any, have landed on my lap with a user asking me to help import/decode the document. The ones that I have encountered (through support contracts I have outstanding) are from clients who are solely interested in extracting the text contents from a document for archival and searching purposes, so I haven't really had much motivation to improve the support for these formats."

Lachowicz was asked about the processes involved in filling the gaps in file format documentation and the difficulties encountered. *"Some guesswork, examination of files/"reverse engineering", hex editors, and just following some general patterns that MS engineers seemed to follow. Of course it'd be easier with complete and total knowledge of the format, access to the engineers, sample source code, etc... But I didn't think that was realistic. Even most OSS projects usually don't live up to that level of documentation and utility."*

Lachowicz was asked about the complexity of file format documentation from sources like Microsoft. *"I don't believe the SPECs were deliberately made difficult to follow. The file format itself is enormously large. IT people generally aren't the best*

spec writers, documentation people generally miss out on or misrepresent important pieces of information, no matter how many peer revisions you go through. I'd always expect MS to have the best knowledge of their own format, as I do of the AbiWord format, for instance. They have access to the original developers, specs, design documents, etc... And I don't think that it's reasonable or fair for them to share all of those things with the community either, though it would be 'nice' if they did so."

Lachowicz was asked how hard it would be to provide support to read in a format like Microsoft Word without any file format documentation. *"Extremely difficult. Not impossible, but very close to."*

5.5 Commercial Developers and Software

There are many commercial document conversion and viewing applications which may be of use to the digital preservation community. In most cases, this software is not Open Source but might be suitable for some preservation purposes.

A range of developers and software tools are described below. Detail and discussion is also made on two developers with contrasting policies towards the release of file format information.

Convert Doc (Softinterface Inc) [32]
http://www.softinterface.com/

'Convert Doc' is described as *"a simple to use, yet sophisticated document conversion utility. If you need to convert thousands of files with a variety of file types located within many folders in a short period of time, this is the tool. Especially if you require complicated conversion jobs done on a regular basis."* [32]

The software provides the facility to migrate between a small number of text based formats in controlled batch processes. Its also possible to utilise the import/export features of Word to perform the migration.

Quick View Plus (Stellent) [33]
http://www.stellent.com/

Stellent provides technology for Content Management systems as well as migration and viewing software. Stellent are described by independent analysts CMS Watch (http://www.cmswatch.com/News/Article/?138) as trying *"…to have a can opener for every conceivable content format…"* [33]. Quick View Plus does indeed support over 225 files types, allowing the user to view, search and print. The Public Record Office are beginning to make use of Quick View Plus to enable rendering of preserved documents.

As the following list of supported word processing formats indicates, a lot of effort has been invested in the development of this tool.

Generic Text

ANSI Text 7 & 8 bit
ASCII Text 7 & 8 bit
HTML Versions through 3.0
IBM FFT All versions
IBM Revisable Form Text All versions
Microsoft Rich Text Format (RTF) All versions
Unicode Text All versions
WML Version 1.2

DOS Word Processors

DEC WPS Plus (DX) Versions through 4.0
DEC WPS Plus (WPL) Versions through 4.1
DisplayWrite 2 & 3 (TXT) All versions
DisplayWrite 4 & 5 Versions through Release 2.0
Enable Versions 3.0, 4.0 and 4.5
First Choice Versions through 3.0
Framework Version 3.0
IBM Writing Assistant Version 1.01
Lotus Manuscript Version 2.0
MASS11 Versions through 8.0
Microsoft Word Versions through 6.0
Microsoft Works Versions through 2.0
MultiMate Versions through 4.0
Navy DIF All versions
Nota Bene Version 3.0
Office Writer Versions 4.0 - 6.0
PC-File Letter Versions through 5.0
PC-File+ Letter Versions through 3.0
PFS:Write Versions A, B and C
Professional Write Versions through 2.1
Q&A Version 2.0
Samna Word Versions through Samna Word IV+
SmartWare II Version 1.02
Sprint Versions through 1.0
Total Word Version 1.2
Volkswriter 3 & 4 Versions through 1.0
Wang PC (IWP) Versions through 2.6
WordMARC Versions through Composer Plus
WordPerfect Versions through 6.1
WordStar Versions through 7.0
WordStar 2000 Versions through 3.0
XyWrite Versions through III Plus

Windows Word Processors

Adobe FrameMaker (MIF) Version 6.0, text only
AMI/AMI Professional Versions through 3.1

Corel/Novell WordPerfect for Windows Versions through 10.0
JustSystems Ichitaro Versions 5.0, 6.0, 8.0, 9.0, 10.0
JustWrite Versions through 3.0
Legacy Versions through 1.1
Lotus Word Pro Versions 96, 97 and Millennium Edition 9.6
Microsoft Windows Works Versions through 4.0
Microsoft Windows Write Versions through 3.0
Microsoft Word for Windows Versions through 2002
Microsoft WordPad All versions
Novell Perfect Works Version 2.0
Professional Write Plus Version 1.0
Q&A Write for Windows Version 3.0
StarOffice Writer Version 5.2
WordStar for Windows Version 1.0
Macintosh Word Processors
MacWrite II Version 1.1
Microsoft Word Versions 4.0 through 98, 2001
Microsoft Works Versions through 2.0
Novell WordPerfect Versions 1.02 through 3.0

This is a comprehensive list of textual file formats, but it should be noted that many versions of formats are not supported, particularly earlier versions. This gives an indication of the scale of the task and the quantity of file format information required. An interesting comparison can be made with the list of documentation on textual file formats on the File Format Encyclopaedia site (see above). A great deal more formats will need to be supported by preservation tools than the documentation is currently available for.

It seems unlikely that Stellent will be interested in sharing file format information with the preservation community as that is the foundation for the quality and range of support of one of their key products. For similar reasons it also seems unlikely that source code for tools like Quick View Plus would be freely released. This raises the danger of dependence on supply from a commercial developer of a long term preservation tool. The Quick View product appears to have changed hands at least once in recent years.

Quick View Plus does at least indicate that the task of supporting many file formats is not impossible and that documentation can be sourced. Dialogue with Stellent as to its sources of file format information may prove worthwhile. Obviously Stellent may want to protect its sources which are of commercial value.

The existence of applications like Quick View Plus may be a useful issue to raise with commercial application developers. If there are already products available that render previously protected proprietary file formats, what do the application developers have to lose by releasing their file format documentation to the preservation community?

Conversions Plus [34]
http://www.dataviz.com/products/conversionsplus/

Conversions Plus is a commercial migration tool, providing migration paths between many different formats. The web site describes the product as follows *"Conversions Plus has become the industry standard for file translation due to the quality of the finished product. For more than 15 years, DataViz customers have relied on quick, stable and quality file translations to keep them compatible."* [34]

"What Comes Through In Translation?" [35] provides an interesting breakdown of the supported formats and describes features of the original documents that will be preserved. Migrations are provided at two levels of quality. For example in textual documents, Level 1 support includes *"bold, italics, underlines, subscript, superscript, shadow, outline, strikeout, all caps, small caps, redline, font size"* and level 2 includes *"color, fonts, hidden text, double underline and styles"*. The following textual formats are supported at levels 1 and 2:

- *AmiPro*
- *AppleWorks GS WP*
- *AppleWorks WP*
- *ClarisWorks Mac & PC*
- *DCA-RFT*
- *FrameMaker MIF*
- *MacWrite II*
- *MacWritePro*
- *MultiMate*
- *MS Works Mac & PC*
- *RTF (Rich Text Format)*
- *Word Mac& Word PC*
- *Word Perfect Mac & PC*
- *WordPerfect Works*

And these formats are only supported at level 1:

- *MS Works Mac 2*
- *WordStar 7 and below*

The listing is clearly not very specific about which versions of the file formats are supported.

Adobe [36]
http://www.adobe.co.uk/

Adobe has an open policy to the use and support of its file formats. Detailed file format documentation is freely available from its web site. Many developers attempt to protect their market share by keeping file format documentation secret and hence preventing or at least hindering the ability of other software to read their file formats. Adobe have taken the opposite approach by releasing documentation and even assisting third party developers in supporting their formats. This has so far proven

very successful for Adobe, with the PDF format in particular seeing very widespread take up. Adobe's PDF format is discussed in many other sections of this report.

Adobe Illustrator is one of the formats used in the example at the start of this report. Adobe's documentation of the Illustrator file format can be freely downloaded from their web site. It provides a very interesting example with regard to the design, clarity and completeness of file format documentation.

Adobe's Illustrator file format document provides one specification for the following versions of the Illustrator file format:

- *Adobe Illustrator 1.0/1.1*
- *Adobe Illustrator 88*
- *Adobe Illustrator 3.0/3.2*
- *Adobe Illustrator 4.0*
- *Adobe Illustrator 5.0/5.5*
- *Adobe Illustrator 5.x, Japanese Edition*
- *Adobe Illustrator 6.0*
- *Adobe Illustrator 7.0*
- *Adobe Illustrator EPS (Encapsulated PostScript)*

The specification uses glyphs to highlight particular sections of information which only apply to specific versions of the format. The specification states that *"…most remarks that apply to early versions of Adobe Illustrator files also apply to later versions, and later versions exhibit the added complexity of their more advanced feature sets."* This leaves a degree of ambiguity in the documentation, further complicated by the overlap of the Illustrator format with the Postscript format.

Adobe's file format documentation can be downloaded from their developer site (open to all) at http://partners.adobe.com/asn/techresources.jsp

Macromedia [37]
http://www.macromedia.com/uk/

An extensive search of file format web sites revealed no information on Macromedia file formats. It is evident from the LA Times example at the start of this report that there is already an impending danger of the loss of data stored in early Macromedia Freehand formats. Contact was made with Macromedia to discover if they were willing to release any file format documentation, and if not, what the reasons were behind this policy. Informal discussions were made with a number of Macromedia employees and a formal response was also provided, as follows.

"We understand the concern and comments that the customer is making regarding the lack of compatibility for the early versions of FreeHand. However, due to the complexities of the file format and the changes that have taken place over the years it is virtually impossible to import an older file successfully into the later version without shifts and loss of data. At this point there are no plans to open the FreeHand file format to the public as open source, so unfortunately the answer to the customer is no, we cannot disclose the requested technical info describing the format."

This response raises a number of interesting points. If the developers themselves feel that migration from older versions of a format to newer versions of a format is *"virtually impossible"*, it seems likely that the task facing the preservation community is considerable. This is especially so if file format information cannot be obtained. However, it seems unlikely that migration cannot be performed at all. Without access to the file format documentation this can really only be considered as speculation.

Many developers do not release file format information in order to maintain control of their formats and to prevent other software from supporting them. It is possible that this represents some of Macromedia's motivation not to release information on its file formats. It may be possible to persuade companies that have thus far been unwilling to release documentation that the release of *dated* documentation would help the digital preservation community greatly, without compromising commercial interests. In the Macromedia example, the release of file format documentation on Freehand versions 1 to 5 would be very unlikely to impact on Macromedia's control of the Freehand format and may in fact make Freehand a more attractive software package to many users. This has certainly been the case with Adobe and the general uptake of formats like PDF.

Microsoft [38]
http://www.microsoft.com/uk/

Microsoft provides information for developers through its MSDN web site [39] and developer CDs. Some documentation aimed primarily at migrating data to Microsoft formats is provided, as well as a specification for Rich Text Format (RTF). Although the MSDN site contains a great deal of valuable documentation for developers it does not contain specifications for Office file formats.

The Wvware site [28] describes the release by Microsoft of some file format specifications in 1998. Some of these have since found there way onto web sites like Wotzit [6] and FFE [8] (see above) *"The MS Office file formats (Word, Excel, Powerpoint, Office inder and Office Drawing) were all made freely available from the MS msdn website in 1998. Since then they have been removed, but MS made cd's available of their website to developers that registered to receive them. These cd's are commonly available. The particular cd that the specifications were made available on is the July 1998 edition. CD Number 2 of the three part set. The specs that were made available were the office 97 specifications. Not the previous versions. The specs are quite hard to read, and often incomplete. Some fields are wrong, and some information is not fully correct, but theres nothing better available."*

Other sources

In "A blueprint for Representation Information in the OAIS model" [40] Holdsworth and Sergeant suggest that anti-virus software developers have reverse engineered proprietary formats. Anti-virus developers might be willing to share technical documentation with the preservation community. Commercial advantage may work

against this possibility, but certainly with regard to older formats, cooperation could well be possible.

5.6 Published sources

Six books were examined:

- 'More File Formats for Popular PC Software', Jeff Walden, 1987 [41]
- 'File Formats', Allen G. Taylor, 1992 [42]
- 'Inside Windows File Formats', Tom Swan, 1993 [43]
- 'Graphics File Formats, 2nd Edition', David C. Kay & John R. Levine, 1995 [44]
- 'Internet File Formats', Tim Kientzle, 1995 [45]
- 'The File Formats Handbook', Günter Born, 1995 [46]

The books appear to be aimed at budding programmers of moderate ability. As a consequence some of the information is simplified, summarised or incomplete. Most of what is provided is presented in a more clear and understandable way than technical specifications.

It is interesting to note that various searches failed to find any file format related books to be in print. The most recent books in this field were first published in 1995. Is the lack of new publications in this area part of a trend?

'Inside Windows File Formats' [43] describes the WMF (Windows Meta File) graphic format. Phil Mellor gained experience of using the WMF format during development of a vector graphic Migration on Request tool, and was asked to assess this information for accuracy and usefulness. *"It gives basic details and a useful commentary on the structure and concepts of a metafile - such as the contents of the header blocks and how internal objects are created, selected and deleted. Unfortunately, very little information is given about the GDI records contained in WMFs that describe the graphical functions that generate the image. A table of identifier constants is supplied, but there is no indication of the different parameters that should be supplied with each record. There is a tutorial on using the API provided by the <windows.h> header file, and the preface to the book indicates that such header files can be referred to for details on the record structures. This makes interpreting a WMF file fairly difficult without the relevant header files and libraries; on its own the book provides little more than a first step to reverse engineering."*

'Graphics File Formats' [44] goes into greater detail. *"The parameters of many common GDI record types are provided, along with useful information about how co-ordinates and colours are represented. This is much more useful for a programmer trying to interpret WMFs from scratch rather than using an existing API and library. Annotated examples of WMFs were also provided which could be helpful as they explicitly show how the various file headers, structures and their parameters fit together." However, says Mellor, "'Inside Windows File Formats' [43] is still*

worthwhile as it offers background material with greater clarity."

It can been seen from this example that this variety of book can often aid understanding but not necessarily provide complete definition of a file format. If the format is fairly simple and doesn't have too many variations and elements, it is more likely to be comprehensively covered. The specifications for PostScript or PDF run to many hundreds of pages so it is unlikely that a book that tries to cover ten formats with fewer pages in total will struggle to give more than the basics or an overview.

Another potential problem is that when omissions occur there is either no information of the type of material that has been omitted or no indication that this is the case at all. This makes it harder to be confident that a file reader would be capable of working with many real world files. The following quotes demonstrate a few of the levels of detail offered by 'Graphics File Formats' [44]:

"Fully detailing the PostScript language requires far more space than is available here. You should refer to the PostScript Language Reference Manual to implement a full PostScript writer or reader. It is our intention here to provide only the information you would generally need to create and read a simple file."

"In this book, we will deal only with communicating standard geometric and graphic information using a single DXF data file ... The DXF characteristics discussed here roughly follow the 80/20 rule; that is, 80 percent of the information is used only 20 percent of the time! ... Much of the following information is abstracted from the AutoCAD Reference Manual"

"The PCX format is relatively simple to write, but tricky to read ... Accordingly, the following description focuses on the worst case: reading a PCX file of indeterminate characteristics and age."

"The CompuServe GIF specification is sufficiently clear, so we simply reproduced it ... by permission of CompuServe Incorporated."

It is not always clear where the information in the books is derived from. For example, the 'File Formats' [42] book thanks various companies for providing information but it is not clear how much of the book's content was sourced from them; 'More File Formats for Popular PC Software' [41] credits sources by name which would (in at least the short term) assist in attributing information. 'Inside Windows File Formats' [43] does not state where information was gathered - it may have been from Microsoft, some other source, or even reverse engineered. Each chapter of 'Graphics File Formats' [44] includes at least one reference, typically to a reputable original source such as the format's specification or an article by its creator. 'More File Formats' [41] states that the information was provided by the format's manufacturer.

Without some idea of the information's source, it can be hard to trust its authenticity and reliability. This makes it harder to correct mistakes by cross reference. Even with a reputable source there can be problems:

"Throughout the year, Lotus staff repeatedly referred us to their FTP archive that contains 1-2-3 .wk1 format specifications. These specifications were indirectly certified by Walden (1986), who describes the specification in detail and provides a sample .wk1 file analyzed byte by byte. Unfortunately, these specifications are incomplete and describe the .wks file format, the format of 1-2-3 release 1A. We were surprised that Walden made such an oversight, but Wotsit's Format Web site (Oliver 1999) and the comp.apps.spreadsheets FAQ (1999) repeat the error. It is clear that neither the professionals nor the amateurs recognized the mistake." [21]

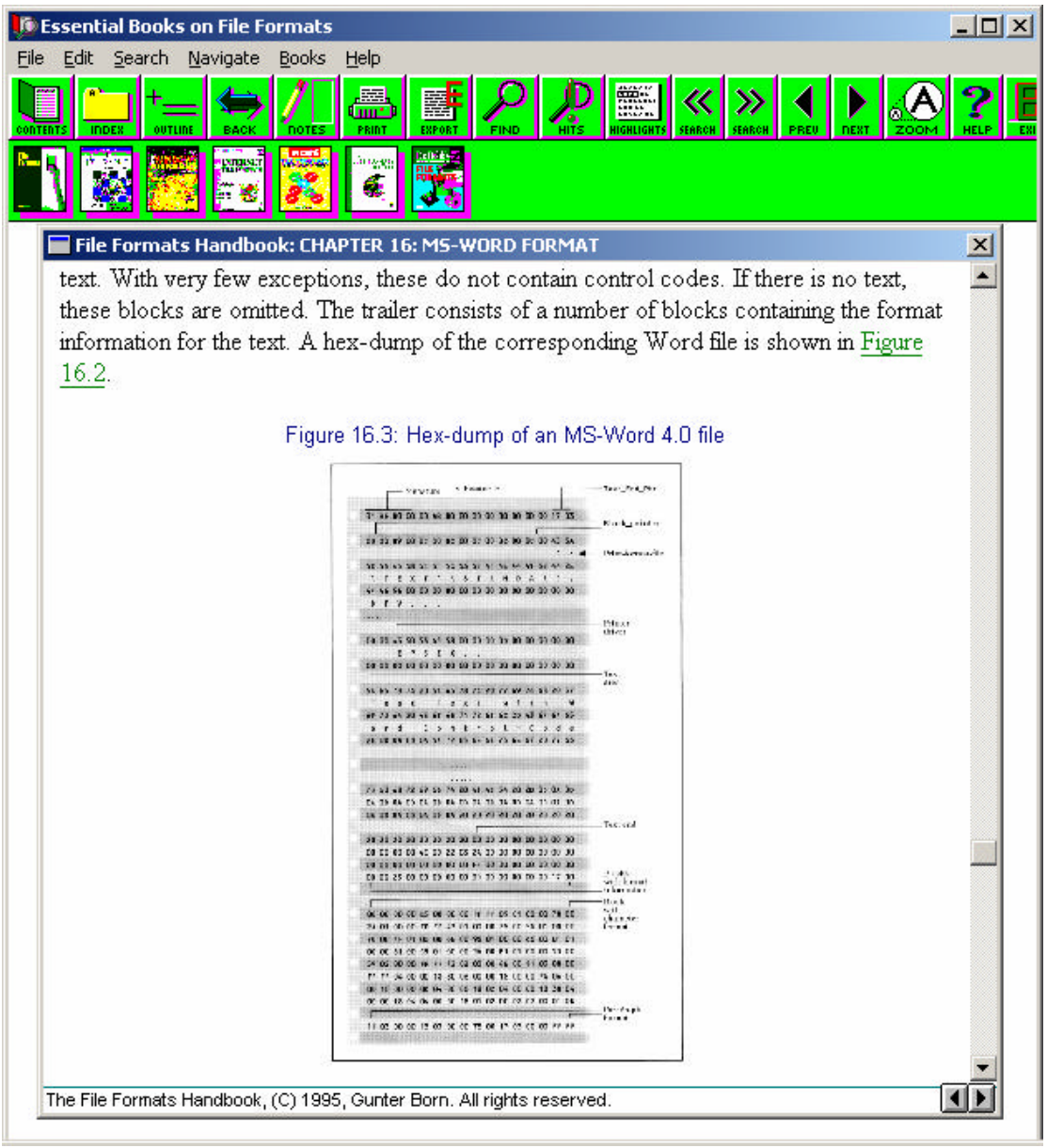
The books did not always specify which version of the file format was being described although an educated guess could usually be made from the date of publication.

In summary, these books offer good background material. Depending on the complexity of the format in question, they can provide anything from a primer for using the format, a guide to the handling APIs, an overview of the format's structure, or a reasonably detailed description of the format.

A compilation CD containing the text of six books is available. This gives the advantage of searching the books very quickly for key words. Unfortunately the books are presented using a proprietary hypertext system (called HyperReader) that runs on the Windows platform. The text can be exported as simple ASCII although this requires the ability to run the original software. There is no separate tool available to the average user to migrate the HyperReader documents into a modern format such as HTML.

The supplied HyperReader application for reading the documents is fairly old (early-mid 1990s), and failed to install on a modern PC running Windows XP. This unfortunately demonstrates how the electronic books will be harder to access in the future. The software runs on earlier operating systems such as Windows 2000, albeit with some minor display problems when scrolling the display.

The electronic books have their disadvantages - some of the tables are rendered as low resolution bitmaps which makes it difficult to read the small text (see below). In "The File Formats Handbook" [46] this is particularly bad - hex dumps of sample files are simply unreadable. Also the reader application has an unfriendly and slightly unfamiliar user interface compared with modern hypertext systems.



5.7 File format identification

Identifying the type of file format of a digital object will be a crucial function on ingest to a digital repository. Automatic identification will be crucial in keeping costs at realistic levels. Understanding the structure of different file formats is obviously crucial to enabling this facility. There are many applications and OS commands that provide file identification, of which the following are a sample.

"File" command for Windows [47]
http://sourceforge.net/project/shownotes.php?release_id=98302

"File tests each argument in an attempt to classify it. There are three sets of tests, performed in this order: filesystem tests, magic number tests, and language tests. The

first test that succeeds causes the file type to be printed. The type printed will usually contain one of the words text (the file contains only printing characters and a few common control characters and is probably safe to read on an ASCII terminal), executable (the file contains the result of compiling a program in a form understandable to some UNIX kernel or another), or data meaning anything else (data is usually `binary' or non-printable). Exceptions are well-known file formats (core files, tar archives) that are known to contain binary data." [47]

More about the "file" command [48]
http://www.cinq.com/linux/tips/file_command.html

"By performing a series of tests, the file command analyzes a file's contents. It also will attempt to determine whether a text file contains C source code, a shell script, or simply ASCII text. If it's a binary executable file, it attempts to determine the format (e.g., an ELF 32-bit MSB executable SPARC Version 1, dynamically linked, stripped file). The file command becomes especially useful when examining an unfamiliar directory and you want to get a quick handle on what files are likely configuration and data files, which are programs and which are scripts." [48]

TypeFind [49]
http://web.archive.org/web/20011016063348/http://www.geocities.com/SiliconValley/Park/4119/typefind.htm

"TypeFind is used to analyse the contents of a file and then take a guess at its filetype from a recognised list of over 200 file formats. This is very handy if a file doesn't have any filetype information for itself, which often happens with Acorn files sent over the world wide web or email, downloaded from bulletin boards and recovered from corrupted disks." [49]

MMagic - Perl [50]
http://search.cpan.org/author/KNOK/File-MMagic/MMagic.pm

"The File::MMagic module can be used to determine the mime type of a file. It uses all kinds of cunning to do this. Firstly it uses a database of "magic" numbers to look at the first few bytes for telltale signs - for example GIF files start with "GIF" and flash files start with "FWS". If that fails - for example html files don't start with anything special - then the module can use extra regular expression techniques to check both the filename and the contents of the file for give away signs that distinguish them." [51]

Research and Development

Various institutions are developing technologies for file identification. The Public Record Office's PRONOM [52] system will feature a simple file format identification based on the recognition of magic numbers in the headers of digital objects. John

Mark Ockerbloom is implementing a more advanced system as part of the development of the Typed Object Model [53]. Ockerbloom explains *"The format identification system that I have allows one to prioritize rules, and also note whether the rule is necessary, sufficient, or both. That way, I can have more reliable rules, or ones that are easier to evaluate, checked before less useful ones.*

I would emphasize the need to include source notations, so one can evaluate the trustworthiness of the rule and to revise them as needed. Note that in practice, a rule set often needs to be 'tuned' for the particular set of formats one is testing for; you can't just rely on eternal, unchanging rules associated with one format, and not take other formats into account."

5.8 Software documentation

The CAMiLEON project [54] highlighted the key role emulation will play in future digital preservation work. The project's proof of concept work in emulating the BBC Domesday system, illustrated the importance of access to both technical and user documentation. Technical documentation is essential to enable the emulation of obsolete computer systems. User documentation is required to assist the user in operating original software running under emulation.

User documentation

In the UK, software user documentation has not fallen within the remit of Legal Deposit. No single encompassing source for documentation of this kind was found. The Public Record Office (The National Archives) [55] has identified the need for collecting old versions of software and their respective user documentation to aid its digital preservation tasks.. Various web sites offer access to user documentation for many different applications. In most of these cases, the company which developed the software in question no longer exists, but use of the software has been revived through the growing open source emulation community.

The BBC Documentation Project [56] is a good example of the material available on the web. Hundreds of printed sources of documentation on BBC Micro software and hardware have been digitised by volunteers before being placed on the web site for reference by users and developers. In some cases OCR has been used to produce searchable documentation, stored primarily in RTF format. Many of the documents are now extremely rare.

The web site highlights the incredibly valuable work performed by enthusiasts which is of great use to the digital preservation community. In the case of the BBC Documentation Project, the web site has not been updated for some time and is dependent for its survival on one individual.

Outside of the UK, French legal deposit covers software documentation supplied with software packages and some documentation is held by the Library of Congress in the USA. All software is not covered by legal deposit although many developers register and deposit their documentation as a way to establish copyright.

Technical documentation

Most manufacturers maintain their own archives of technical documentation which can be invaluable sources of information for emulation developers. Where manufacturers or developers have survived the commercial world for many years, as is the case with IBM for example, this information will still be available. Many companies did not survive and their documentation has often only been preserved in the hands of former employees or enthusiasts.

The Science Museum Documentation Centre has a considerable collection of technical documentation which can be found on the Computer Conservation Society's web pages [57]. This includes many ICL documents which have been maintained since the demise of the company. Another archive is currently being established at Bletchley Park.

Mauriton [58] is a supplier of operating manuals and service guides which can be of great use, especially when dealing with peripherals.

Experiences at ULCC

Kevin Ashley was asked about sources of documentation used at ULCC [59]. *"We use a mix of pooled knowledge (some of us have been around longer than we would admit to), wotsit.org for more recent stuff, newsgroups such as alt.folklore.computers (or sometimes more specific ones), reverse-engineering and code-cracking in some cases, blind guesses, and manufacturer libraries/archives (IBM have helped in this respect)."*

Ashley was asked if preservation work had ever been halted due to a lack of documentation. *"Not often. Only admitted defeat once so far, and that was only because the supplier definitely had the format spec but wanted to charge the owner of the data umpteen gazillions to convert it to a more useful format. We weren't willing to do that work for free when we knew they shouldn't have been charging at all as it was built into the contract that they should have done this.*

That said, sometimes we've had to use guesswork, data dumps and comparisons with printouts to do it. There are times we've had no documentation at all."

Experiences on CAMiLEON [60]
http://www.si.umich.edu/CAMILEON/reports/cingahd.html

Work conducted by the CAMiLEON project and the Computer Conservation Society to emulate an ICL1700 computer from the 1970s provided valuable insights into the knowledge required to write emulators.

David Holdsworth and Delwyn Holdroyd developed the ICL1700 emulator from a combination of documentation and personal experience of the original system.

Holdsworth describes *"The George3 operating system which runs under our emulator was written in assembler by a team of programmers. As a result it seems to use every quirk of the machine's order code at some point. A final break-through into reliable operation came when we finally implemented a property of the overflow register that was not hinted at in the summary chart, and was detailed once in a thick four-volume manual. It seems likely that such a property might escape the specification process."* Here Holdsworth argues against the concept of Encapsulation, where a specification of data or of a system is preserved with a digital object. In theory this provides the information required to construct a rendering tool at a later date. Many bugs and features of both hardware and software are often left undocumented. The only test of and encapsulated specification is at the point it is used to implement a rendering tool. The risk of missing vital information in the specification seems to invalidate this approach.

Holdsworth argues for the value of source code in the development of both emulation and migration tools. In this example, source code from the original system aided the development of the rendering tool which preserved it. *"The source text of George3 was an invaluable reference from time to time. Some of the later features of the system's interfaces were not in the main stream manuals, although they may have featured in software notices. The thought of reading through many hundreds of these was sufficient disincentive to make inspection of the source code a more fruitful way to investigate mysteries. One particular feature of the interface to the communications processor was only revealed by a comment in the source code, after which dim recollection of 25 year-old knowledge was sufficient."*

"During the early stages of the emulation work, there still existed a single live installation of George3. We took the precaution of getting this system to produce its diagnostic memory dump, so that we had an example of a real system in operation. Reference to this did occasionally help us to clarify aspects of interfaces whose documentation assumed knowledge that we no longer possessed." Access to the original system is often vital in producing new rendering tools. The issue of timeliness of preservation is discussed in more detail in the CAMiLEON Project final reports [61]

Experience from the CAMiLEON project indicates that where official preservation of documentation fails, individuals often attempt to rescue what they can. CAMiLEON's work in emulating BBC Domesday was made easier by the contribution of technical documentation by individuals involved in the Domesday project. In many cases individuals rescued documentation that otherwise would have been thrown away. Over time, this information will be lost as attics and basements are "tidied up". One possibility would be to organise an appeal for donations to what would become a technical documentation library.

6.0 Initiatives to collate and deliver file format and representation information

A number of projects are investigating or developing systems which provide repositories of file format and representation information for use in digital preservation.

PRONOM [52]
http://www.pro.gov.uk/about/preservation/digital/pronom.htm

The Public Record Office is developing a database system that *"…stores and provides information about file formats and the application software needed to open them. File format information is vital to digital preservation. Without it, a unit of data is merely an unintelligible stream of ones and zeros. The task of preserving digital objects requires a reliable, sustained repository of file format information."* [52] PRONOM can be obtained for use outside of the PRO, but will be supplied as an empty system only, without the populated file format information collected by the PRO. It is hoped that version 3 of the system will be available for access over the internet.

File Format Registry [62]
http://hul.harvard.edu/~stephen/Format_Registry.doc

Harvard University and the Massachusetts Institute of Technology are leading an initiative to establish a registry of file format information. The initiative is currently at an early stage but has already seen international interest and contribution from a range of institutions and organisations facing the same digital preservation problems. It has been identified at an early stage that there will be significant advantage in pooling experience on file formats and sharing file format information. It is likely that the central registry will be based in North America but will involve global collaboration.

Representation Networks [63]
http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

For digital objects to be preserved for a long time, a knowledge of the format and tools that work with that format also need to be preserved. Otherwise the ability to extract the information from those digital objects is lost. Representation Information is the section of metadata where this knowledge is preserved. Each component of the Representation Information will also need preserving along with a knowledge of how to use that Representation Information. Combining Representation Information into Representation Networks enables common format knowledge and tools to be shared between different digital objects without the need for expensive replication.

As new Representation Information becomes available (for example a tool that allows a digital object to be embedded in a multimedia - document, or a Migration on Request module enables transformation into a new format) this can easily be added on

to the existing Representation Networks. At the simplest level, Representation Networks provide a manageable way of associating digital objects with their file format specifications and with tools designed to work with them. Representation Networks also provide the backbone for maintaining rendering software and access software, and detecting obsolescent platforms: ensuring that the digital objects continue to be meaningful in the long-term.

The Representation and Rendering project has developed a Representation Network for some documents describing the rescue of the BBC Domesday. This also includes preserving the software that moves the information on the 12 inch videodiscs into a media neutral form. The process of building the Representation Network is documented, and research into maintaining these is ongoing.

Further information can be found in the "The Cedars guide to Cedars Distributed Archiving Prototype" [64] and a Representation Network object model and rough example can be found in the OAIS red book [65].

Digital Curation Centre (JISC)
http://www.jisc.ac.uk/index.cfm?name=pres_curation_draft

JISC has recognised the pressing need for the establishment of a Digital Curation Centre. At the time of writing JISC has released a draft ITT which describes the likely functions of the Centre. *"…the Digital Curation Centre is not intended to be an archive or repository itself. It is developing a set of shared services such as preservation watch, registries and tools for such repositories. It is also acting as a catalyst supporting the vision of a distributed network of curation consisting of many different funders and types of repository, proposed in the JISC Continuing Access and Digital Preservation Strategy 2002-5."* [67]

7.0 Conclusions

The wide range of sources of file format information described in this report is only a sample of what is available but points the way forward for future collection of this material. A priority must be given to capturing and preserving documentation currently available on web sites which have no guarantee of permanence.

The experiences of the CAMiLEON project suggest that emulation will play a significant role in the preservation of digital materials, particularly where digital objects contain interactive or executable elements. Emulation may also be crucial as a backup to the migration strategies used to preserve "passive" digital documents. Verification of the success of migration will be possible by using emulation to render a digital object in its original environment. And in some cases migration of passive documents may not be possible where appropriate file format documentation cannot be obtained. It seems almost certain that some developers will always want to protect their commercial interests with secrecy about the composition of their file formats. In this situation, emulation may not be the ideal method of rendering for the user but it might be the only method available. Collating a repository of software documentation will be crucial to enable emulation strategies to be pursued in the future by ensuring that a backup to migration strategies is available where necessary.

Existing sources of file format information fall far short of what is required to successfully tackle the problems of data obsolescence, but they provide a significant starting point for the preservation community. Developing collections of file format documentation and fostering links with commercial developers will be essential in expanding the range of understandable formats. Developers who have not released any file format documentation will need to be encouraged that at the very least, older documentation can be released without necessarily compromising current formats. In many cases it will be in the developers favour to do this, as companies like Adobe have demonstrated. In some cases, commercial rendering tools already support protected proprietary formats, so it can be argued that developers have nothing left to gain by keeping their documentation to themselves. High profile lobbying from organisations like the Digital Preservation Coalition will be crucial in this area.

The accuracy of the majority of available file format information is mediocre at best. An analysis of a specification's accuracy can be made after its use in the implementation of a rendering tool that decodes the format in question. Evidence from rendering tool developers reveals many inaccuracies and omissions from most the of the file format information available. Even the best file format information is not perfect. This does not mean that the long term preservation of many kinds of digital objects will be impossible to achieve. Evidence, again from the developers of rendering tools, suggests that with no file format documentation available the effort required to support a specific file format would be excessive. But with reasonable (if not perfect) documentation as a starting point, the effort required to fill in the blanks is not unrealistic. By constructing files and observing their behaviour in the original application software, it is possible to ascertain the finer points of a file format specification even if they aren't described in the documentation.

If action is taken to find, negotiate for, obtain and secure file format information for use by the digital preservation community, it is predicted that rendering solutions could realistically be provided for the majority of digital objects that need to be preserved. Undoubtedly there will always be some formats for which preservation by migration type strategies will not be possible due to lack of knowledge about the format. In the majority of these cases emulation should be able to provide an alternative preservation solution. In this case, technical and user documentation will be required to inform the emulation development process.

8.0 Recommendations

This report makes the following recommendations to JISC and its strategy for the long term preservation of digital objects.

8.1 Urgent recommendations

As has been illustrated, there is a pressing need to take action to ensure the information required to perform digital preservation activities is not lost. It is suggested that the following recommendations be carried out as soon as possible.

8.11 Collection and preservation of available file format documentation.

There is a pressing need to capture file format documentation currently available on the web. This information should be collected and preserved in a repository.

8.12 Collection and preservation of technical hardware and software documentation

The Science Museum in conjunction with the Computer Conservation Society are currently working to preserve and provide access to a substantial quantity of technical documentation. This report recommends that assistance be given to these organisations to collect and preserve technical documentation at risk of loss.

8.2 Essential recommendations

Essential recommendations are seen as being crucial to a national strategy for the long term preservation of digital information.

8.21 Establishment of a system for Representation Information

Establishment of a repository of file format information, ideally as part of an OAIS Representation Network. This would record information on different file formats, rendering tools and the platforms on which those tools run. The system would provide Representation Metadata which could be referenced by individual digital repositories. The Representation Network would include a "technology watch" function to monitor for the obsolescence of rendering tools.

8.22 Collection of existing rendering tools

A great deal of open source rendering tools are available on the web and many commercial applications also provide rendering solutions. It is recommended that the Representation Information system be populated with references to these existing rendering solutions. Where appropriate, open source tools should be collected and placed in a secure repository along with documentation.

8.23Development of new rendering tools

Many important digital objects are already close to obsolescence. While existing sources of rendering tools will prove to be useful to the preservation community, they will not provide solutions to the preservation of all digital objects. The development of new long term preservation tools and strategies must be continued. These rendering tools should be described and monitored within appropriate Representation Networks.

8.24 Dialogue with commercial application developers

There are many file formats for which documentation is not available. The preservation community must open a dialogue with commercial application developers and lobby for the release of the required information. It is suggested that a high profile organisation like the Digital Preservation Coalition would be most successful in this role.

8.25 Development of ingest tools

Automated processes for the ingest of digital objects to a repository are seen as essential in providing economical preservation services. It is recommended that effort be devoted to tools which can facilitate this aim. In particular tools for the identification of the format of digital objects will be essential.

8.26 Establishment of testbed for preservation strategies

Independent evaluation of the accuracy, efficiency, long term sustainability and practicality of rendering tools will be crucial for repositories in selecting appropriate preservation strategies. It is recommended that a preservation testbed be established to provide this function. The Testbed Digitale Bewaring project is widely recognised as an exemplar in this field.

8.27 Continued collection of file format and technical documentation

As new formats are created, or old formats are investigated, appropriate file format documentation must be collected and maintained in the Representation Information system.

8.3 Desirable recommendations

Desirable recommendations are seen as being useful contributions to the digital preservation effort but are not priorities in relation to the other recommendations made above.

8.31 Appeal for technical documentation

A public appeal for donations of technical documentation may make up in part for the lack of focus in preserving this kind of material over the last few decades.

9.0 Summary of Recommendations

- Collect and preserve file format documentation currently available on the web that is at risk of loss.

- Establish a system of OAIS Representation Networks to manage and preserve file format documentation, providing a technical metadata solution for digital object repositories.

- Populate the Representation Networks with existing rendering solutions (in particular open source rendering tools).

- Open a dialogue with commercial developers with a view to encouraging the release of file format and software documentation.

- Develop new rendering and ingest tools.

- Establish a testbed to provide independent evaluation of the effectiveness of rendering tools.

- Ensure the ongoing support and maintenance of the Representation Networks system.

10.0 Bibliography

[1] Digital Preservation Coalition http://www.dpconline.org
[2] CAMiLEON, BBC Domesday
http://www.si.umich.edu/CAMILEON/domesday/domesday.html
[3] Testbed Digitale Bewaring project http://www.digitaleduurzaamheid.nl/
[4] "CAMiLEON guide to new digital preservation strategies"
http://www.si.umich.edu/CAMILEON/reports/reports.html
[5] "Migration: Context and Current Status"
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf
[6] Wotzit's Format? http://www.wotsit.org/default.asp
[7] MyFileFormats http://myfileformats.com
[8] File Format Encyclopaedia http://pipin.tmd.ns.ac.yu/extra/fileformat/
[9] Szuper http://www.szuper.biz
[10] The Diffuse Project http://www.diffuse.org
[11] Open Information Interchange initiative
http://www.diffuse.org/oii/en/oiistand.html#oiistand
[12] National Software Reference Library http://www.nsrl.nist.gov/index.html
[13] PDF Zone http://www.pdfzone.com
[14] Archiving and Preserving PDF Files, John Mark Ockerbloom
http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2
[15] Cartagis Survey http://www.cartagis.com/fileformatsurvey.html
[16] Shapelib http://gdal.velocet.ca/projects/shapelib
[17] "The CAMiLEON Guide to New Digital Preservation Strategies"
http://ww.si.umich.edu/CAMILEON/reports/reports.html
[18] PDF-Archive http://www.aiim.org/standards.asp?ID=25013
[19] CAMiLEON, Migration on Request
http://www.si.umich.edu/CAMILEON/reports/mor/index.html
[20] The Florida Center for Library Automation (FCLA)
http://www.fcla.edu/digitalArchive/index.htm
[21] Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger, William H. Walters, Anne R. Kenney, "Risk Management of Digital Information:A File Format Investigation" (June 2000) http://www.clir.org/pubs/reports/pub93/contents.html
[22] Philip Lord, "Public Draft -Extracts from A Survey of Information Technology Vendors", http://www.dpconline.org/graphics/reports/
[23] PDFbox http://www.pdfbox.org/
[24] Xpdf http://www.foolabs.com/xpdf/
[25] Adobe Legal Notices http://partners.adobe.com/asn/developer/legalnotices.jsp
[26] !PDF, http://pdf.iconbar.com/
[27] Open Office http://www.openoffice.org/
[28] wvware http://www.wvware.com/
[29] CRAK http://www.crak.com/
[30] LAOLA http://snake.cs.tu-berlin.de:8081/~schwartz/pmh/index.html
[31] AbiWord http://www.abisource.com/
[32] Softinterface Inc, Convert Doc http://www.softinterface.com/
[33] Stellent, Quick View Plus http://www.stellent.com/
[34] Conversions Plus http://www.dataviz.com/products/conversionsplus/
[35] "What Comes Through In Translation?"
http://www.dataviz.com/products/conversionsplus/cp_details.html#level%20one
[36] Adobe http://www.adobe.co.uk/

[37] Macromedia http://www.macromedia.com/uk/
[38] Microsoft http://www.microsoft.com/uk/
[39] MSDN http://msdn.microsoft.com/library/default.asp
[40] "A blueprint for Representation Information in the OAIS model"
http://www.personal.leeds.ac.uk/~ecldh/cedars/nasa2000/nasa2000.html
[41] Jeff Walden, "More File Formats for Popular PC Software" (1987)
[42] Allen G. Taylor, "File Formats" (1992)
[43] Tom Swan, "Inside Windows File Formats" (1993)
[44] David C. Kay & John R. Levine, "Graphics File Formats, 2nd Edition", (1995)
[45] Tim Kientzle, "Internet File Formats", (1995)
[46] Günter Born, "The File Formats Handbook", (1995)
[47] "File" command for Windows
http://sourceforge.net/project/shownotes.php?release_id=98302
[48] More about the "file" command
http://www.cinq.com/linux/tips/file_command.html
[49] TypeFind
http://web.archive.org/web/20011016063348/http://www.geocities.com/SiliconValley/Park/4119/typefind.htm
[50] MMagic - Perl http://search.cpan.org/author/KNOK/File-MMagic/MMagic.pm
[51] Perl Advent Calendar http://www.perladvent.org/2002/7th/
[52] PRONOM http://www.pro.gov.uk/about/preservation/digital/pronom.htm
[53] Typed Object Model http://tom.library.upenn.edu/
[54] CAMiLEON project http://www.si.umich.edu/CAMILEON/
[55] Public Record Office
http://www.pro.gov.uk/about/preservation/digital/default.htm
[56] BBC Documentation Project http://members.aon.at/~musher/bbc/
[57] Computer Conservation Society's web pages http://www.cs.man.ac.uk/CCS
[58] Mauriton http://www.mauritron.co.uk/mauritron/
[59] ULCC http://www.ulcc.ac.uk/
[60] Experiences on CAMiLEON
http://www.si.umich.edu/CAMILEON/reports/cingahd.html
[61] CAMiLEON final project reports
http://www.si.umich.edu/CAMILEON/reports/reports.html
[62] File Format Registry http://hul.harvard.edu/~stephen/Format_Registry.doc
[63] Representation Networks http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf
[64] "The Cedars guide to Cedars Distributed Archiving Prototype"
http://www.leeds.ac.uk/cedars/guideto/cdap/guidetocdap.pdf
[65] OAIS red book http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
[66] Simon Davis, Digital Preservation Strategy
http://www.naa.gov.au/recordkeeping/noticeboard/Simon_Davis_Digital_Preservation_Strategy.pdf
[67] Discussion paper: draft ITT for a digital curation centre
http://www.jisc.ac.uk/index.cfm?name=pres_curation_draft
[68] E Science curation audit draft report (JISC)
http://www.jisc.ac.uk/index.cfm?name=project_escience