Project Number	IST-2006-033789
Project Title	Planets
Title of Deliverable	Gap analysis: a survey of PA tool provision
Deliverable Number	D3
Contributing Sub-project and Work-package	PA/2
Deliverable Dissemination Level	External Public
Deliverable Nature	Report
Contractual Delivery Date	31th August 2009
Actual Delivery Date	12th October 2009
Author(s)	KB-NL

Abstract

This report looks into the file formats which are archived for the long term by cultural heritage institutions. An inventory of preservation action tools is made, and the use of the Planets Core Registry for the research into gap in tool provision is explored. A preliminary analysis of existing gaps has been made.

Keyword list

Gap analysis, preservation action tools, file format inventory

Contributors

Person	Role	Partner	Contribution
Sara van Bussel	Author	KB-NL	
Frank Houtman	Author	KB-NL	

Document Approval

Person	Role	Partner
Frank Houtman	PA SP Lead	KB-NL
Christen Hedegaard	External Reviewer	KB-DK

Revision History

Issue	Author	Date	Description
1.0	KB-NL	31-08-2007	First iteration
2.0	KB-NL	15-04-2008	Second iteration. Expansion of the first iteration. More sources for file formats archived by institutions. Added analyses of found data.
3.0	KB-NL	7-20-2008	Third version. Expansion of the first two iterations. Added sources of file formats, comparison to pre-existing research on this subject, case study of three file formats, research into preservation action tool inventory and preliminary analysis of gaps.
4.0	KB-NL		Fourth version. First external iteration, gathering of information from previous iteration. List of migration tools, analyses of the Planets Core Registry for gap analysis, general conclusion.

EXECUTIVE SUMMARY

This is the final iteration of the Gap Analysis in Tool Provision. All previous iterations come together in this release. Added to this is more information about preservation action tools and a general conclusion.

A survey was executed to gather information about the file formats archived at cultural and scientific institutions. This resulted in an inventory list of 137 different file formats, submitted by 76 respondents. The inventory shows that while there are a few file formats that are archived in many institutions, it is also true that 76% of all file formats are found in three institutions or less. This inventory is confirmed in a comparison with existing studies. The added value of the file format inventory and this report is that this is the first report where gathering information about the specific file formats archived by cultural and scientific institutions was one of the main goals. This previously unavailable information can not only be used in this report, but might be used in other digital preservation studies.

In a case study of three of these smaller file formats, DAISY (format to make talking books available to users with reading disabilities), FITS (format to store astronomical data) and sheet music formats, it is shown that when a central non-profit consortium of users or developers is behind the development of a file format, there is a good chance that digital preservation issues that arise will be taken care of by this consortium. However, if development is decentralized in a for-profit environment, digital preservation and interoperability are not a priority, leading to issues.

Within Planets, the Planets Core Registry (PCR) is being developed in which information about file formats and preservation action tools (amongst other types of information) will be stored. This registry is still in development, so for this report a list of migration tools was made from submissions by Planets partners. This list contains 57 tools. All but one of the ten most used file formats can be migrated by these tools. The single file format that cannot be migrated is XML, a file format that is used as the output for many migration tools. Upon completion of development of the PCR, this registry can be used to find gaps in tool provision in an automatic way.

Strictly speaking, it can be said that there are no gaps in tool provision. There is nearly always a tool available that can perform a preservation action on an object. However, an institution probably has specific requirements for a tool, concerning operating environment, licensing, and quality. This means that there might not be a tool available for their specific set of requirements, which indicates a gap.

The world of digital objects and digital preservation is constantly evolving. New file formats are adopted, new tools are developed. More information about file formats and tools will be gathered and made available. For these reason the gap analysis should be redone every two to three years to take full advantage of the new information that becomes available.

The results of the analysis not only give important insight in the status of digital preservation, but also in the status of digitization. Even more important a regular gap analysis in tool provision can be used in a justification and indication for new research and/or development of new tools.

TABLE OF CONTENTS

1. Introduction 5
2. File format survey 6
2.1 Introduction 6
2.2 Methodology 6
2.3 Other research 6
2.3.1 Pre-existing research 7
2.3.2 NESTOR study 7
2.3.3 Registry of Open Access Repositories 7
2.4 Analysis of found file formats 8
2.5 Overview 9
3. Preservation action tools 11
3.1 Introduction 11
3.2 Methodology 11
3.3 Migration tools 11
3.4 Emulation tools 11
3.5 Planets Core Registry 12
3.6 Migration tools 12
3.7 Emulation tools 12
3.8 Pathways 13
3.9 Conclusion 13
4. Gap analysis 14
4.1 Introduction 14
4.2 Emulation tools 14
4.3 Migration tools 14
4.4 Plato 14
4.5 Case studies of individual file formats 15
4.5.1 Sheet music file formats 15
4.5.2 FITS 16
4.5.3 DAISY 17
4.5.4 Conclusion 18
4.6 Conclusion 19
5. Conclusion 20
6. Appendices 21

1. Introduction

To preserve digital objects one needs to perform preservation action, in order to perform these actions one needs preservation action tools. To be able to tell what kind of preservation actions should be provided, one needs to know which formats are used for archiving digital information. This in a nutshell is what the Planets Gap Analysis is all about, analyzing which tools do not exist but are needed.

If the Analysis shows no tool for a specific preservation action exists, there definitely is a gap in tool provision and a new tool should be developed. Existing tools can be wrapped and made available within the Planets framework.

This document contains an overview of the work done during the lifetime of the Planets project. In the next chapter the results of the file format survey will be presented and analysed. Of course an inventory of used file formats at cultural heritage institutes says nothing about the availability of preservation action tools. Therefore we need to compare the inventory of file formats with a list of preservation action tools, which will be done in chapter 3.

During the analysis performed on both inventories, it became more and more clear that the search for gaps in tool provision cannot be limited to availability of tools for most occurring file formats. There are many specialized formats that need support from specialized PA tools. This will be researched by several case studies in chapter 4. The report will be closed by drawing several conclusions and some recommendations.

2. File format survey

2.1 Introduction

Within the Planets project, the Preservation Action subproject is responsible for providing the tools that are required to perform preservation actions. In order to do so, existing tools can be wrapped and made available within the Planets framework. If no tool for a certain action exists new tools can be developed. To determine what kind of preservation actions should be provided by the system, and thus which tools should be build or wrapped, one should know what file formats are used for storing the information that needs to be preserved.

This chapter provides an inventory of file formats that are used by various institutions that produce or store information in digital form. To optimize its reliability, this inventory has been extended several times in the last two years. The result is representative and can be used to identify the need for specific preservation actions.

2.2 Methodology

In order to obtain an overview of the file formats that are used for storing cultural and scientific data, the National Library of the Netherlands conducted a survey in which a number of institutions where asked the following questions:

- Which file formats does your institution archive for the long term?
- Do you have any experiences with file formats that appear to be obsolete?
- Which software programs does your institution use for editing/rendering/converting the file formats which are archived for the long term?

The number of questions in this survey was kept very low intentionally, since the goal was to get an indication of used file types and a wide coverage was considered more important than a detailed investigation.

Three surveys were held to gather data for the previous iterations of this report. The first survey was spread amongst the members of Planets in July 2006. The members were asked which file formats were archived for the long term in their repositories and what the percentage of occurrences of these file formats was. Seven surveys were received. In most cases, percentages of occurrences were not given. The second survey was held amongst cultural heritage institutions in the Netherlands, in January 2007. The type of institutions surveyed were museums, libraries, archives, audio-visual archives, universities, data centres and supporting institution. In the second and third iteration the survey was expanded with the results of a questionnaire that was send to institutions in the UK and Denmark. A subsequent survey was undertaken to find respondents in countries that had not received or replied to the survey earlier. The survey was sent out to institutions in Australia, Germany, Finland, France, Norway, Slovenia, Switzerland, Sweden and the United States. The survey was also sent out to institution types that were not adequately represented by the results of the previous surveys.

Based on these results, we created an inventory of file formats that are currently archived for the long term at institutions. This resulting list was also used for several analyses.

2.3 Other research

To complement the results of the surveys other sources have been examined, dealing with the same questions. A search was carried out for sources dealing with digital preservation, and more specifically file formats. We have started at websites of the national coalitions or institutions for digital preservation such as the Digital Curation Centre (DCC) in the UK and the National Digital Information Infrastructure and Preservation Program (NDIIPP) in the USA. By approaching the desktop research in this way we hoped to extend the scope of the file format list to different countries and types of institutions, not present in the surveys.

In 2004, the Digital Curation Centre (DCC) in the UK carried out 6 interviews to assess user requirements for digital curation.[1] Amongst the questions asked were questions about which file formats were used and archived and whether they had problems with these file formats.

The Library of Congress started a website in 2004, the Digital Formats website, to support strategic planning regarding digital content formats. The website identifies and describes formats and identifies whether they are promising for long-term sustainability. With each description they note if they have the format in their collection.

2.3.1 Pre-existing research

In the third iteration there was more information available about file formats that are stored at cultural heritage institutions. This information could answer questions concerning the long term storage of file formats. Because this information was presented in a generic manner, it was not fit for inclusion in the file format inventory, but it did provide more profound information about the subject. Examples are two studies done by NESTOR and ROAR, which are investigated below.

2.3.2 NESTOR study

In 2004, within the NESTOR network for long term digital preservation in Germany, about 1200 German museums answered questions about their digitalisation projects and care for digital objects.[2] One of the questions dealt with the file formats these objects are found in.

Based on their research it was found that text was mostly stored in the DOC format (71.4%), but also as PDF (30.9%). This is different from our findings, where most text is stored as PDF, with DOC a close second. This is also the case if we only look at museums; four museums store text as DOC files, and eight museums store text as PDF.

For images, the study finds that most files are either in the JPEG (64.4%) or TIFF (43%) formats. This is similar to our research; however, in our survey we found that more institutions store TIFF (66%) than JPEG (49%). When looking only at museums, it can be seen that TIFF and JPEG files are found in 78% of all museums.

The last category on which the report focuses are media file formats, which are both audio and visual file formats. The file formats found most are WAV (7.9%), AVI (6.9%), MPEG (1.9%) and MP3 (0.8%). In our research more museums had audio-visual files in their collection. The division was about the same, the file formats that occurred the most were WAV (28%), MP3 (33%), MOV (22%), MPEG (28%) and AVI (17%). Of these five file formats, only MOV is not found in the NESTOR report.

Overall the findings of the NESTOR report are grosso modo similar to the findings of this report. Nearly always the same file formats are found in the museums sector in both reports, however, the ranking of each file format differs slightly. This shows that only relying on the ranking of specific file formats is not enough, because there is not enough reliable data available about the occurrences of file formats in cultural and scientific institutions to make a reliable ranking of importance. This comparison does show that the file formats found for this iteration are also the most archived file formats in German museums, which shows that the file format inventory is useful.

2.3.3 Registry of Open Access Repositories

ROAR is the Registry of Open Access Repositories which contains information about open access e-print archives. They automatically collect information about each repository, including system, number of records that have been uploaded and a description. It is also possible to generate a list of file formats found in the repositories in ROAR. Such a list was part of the first iteration of this report, but was later left out because it does not give any information on the institution type that archives the files, which makes analysis and comparison difficult. However, the list of file formats found in the repositories registered by ROAR can be compared to the inventory made for this

[1] *Digital Curation Centre: Interviews* (2004), found at: http://www.dcc.ac.uk/resource/interviews/ on April 2nd 2008.

[2] D. Withaut, e.a., *Digitalisierung und Erhalt von Digitalisaten in deutschen Museen*, <http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_02.pdf>, accessed on September 1 2008

report. Repositories are not included in the inventory as a separate institution category. Most repositories are held within other institutions, for example libraries and universities, and are listed as such. Individual repositories might form their own category in the future, because they too can be seen as a cultural or scientific institution.

When looking at the file formats at the top of the ROAR list, and grouping the different versions of each file format together, the most popular six file formats are: PDF (65%), HTML (9.4%), JPEG (7.1%), TXT (5.3%), TIFF (3.9%) and XML (1.5%). When compared to the top formats in our Planets inventory, the list is somewhat similar. In the ROAR list PDF and HTML score much higher than in our inventory. This is to be expected however, because of the nature of the repositories in ROAR as e-print archives where the focus is on text rather than images. The above mentioned six file formats are found in the top of the inventory list, the few file formats that are found in the top of the inventory and not in the top of the ROAR list are MP3, WAV, GIF and MPEG. This is likely also a consequence of the set up of ROAR as an e-print archive.

2.4 Analysis of found file formats

In the following some results of analysis based on the list of found file formats (Appendix A) are described.

In the second iteration this list was analysed in several ways. When it was analysed by looking at the occurrences of a given file format several things became clear. Only 22% of the archived file formats were found in four or more institutions, only two file formats were found in over half of all institutions. These two file formats are TIFF and JPEG. This shows that when strategies are based on numbers many file formats and institutions need to be left out.

To prevent this, the file formats in the list have been divided into categories based on the intended content of the file, i.e. audio, video, vector image, plain text etc. This led to a total of 19 categories. Most file formats were found in 6 of these categories: raster images, formatted documents, video, audio, databases and spreadsheets. A high number of file formats in a category does not mean that there is a high number of different institutions with archived file formats from that category. Instead, it might show that there is no main file format in that category that is used most; instead, many file formats are used by only a few institutions. Also, a low number of file formats in a category does not mean they are found in only a few institutions. Here it might be shown that one or two file formats in the category are used by a great deal of institutions, acting as a standard file format for that particular type of file. A closer analysis reveals a tendency towards standardisation; in each category one or two file formats are archived by the bulk of the institutions.

Like the file formats, institutions can be divided into categories. This led to 7 categories: archives, AV archives, libraries, data centres, museum, supporting institutions and universities. Most of the surveyed institutions fit in the three categories of archives, libraries and museums. By sorting the institutions into categories detailed information can be gathered about the importance of certain file formats in specific types of institutions. This does not make a difference when looking at the most archived file formats, overall this is the same in each category. However, further down the list differences become apparent. For example, archives hardly archive MP3 files, while museums do.

The conclusions that can be drawn from this inventory are that TIFF, JPEG and PDF are the three most archived file formats. However, when looking at file format or institution categories, a more detailed analysis can be made of the file formats that are archived. Also, the list reveals there are many file formats that are archived by only one or a few institutions. It is important to keep in mind that this gives no information about the value of these file formats.

In the third iteration and after three major rounds of gathering information, the inventory now lists 137 file formats, submitted by 76 respondents. Out of 137 file formats, 78 (57%) are archived by only one of the 76 institutions, 16 (12%) file formats are archived by two institutions and 10 (7%) are archived by three institutions. This means that 76% of the archived file formats are found three times or less.

In figure 1 below, only file formats that are archived by seven or more institutions (14% of all file formats) are charted to improve readability of the chart. The chart shows us that there are only a few file formats which occur in a large group of institutions.

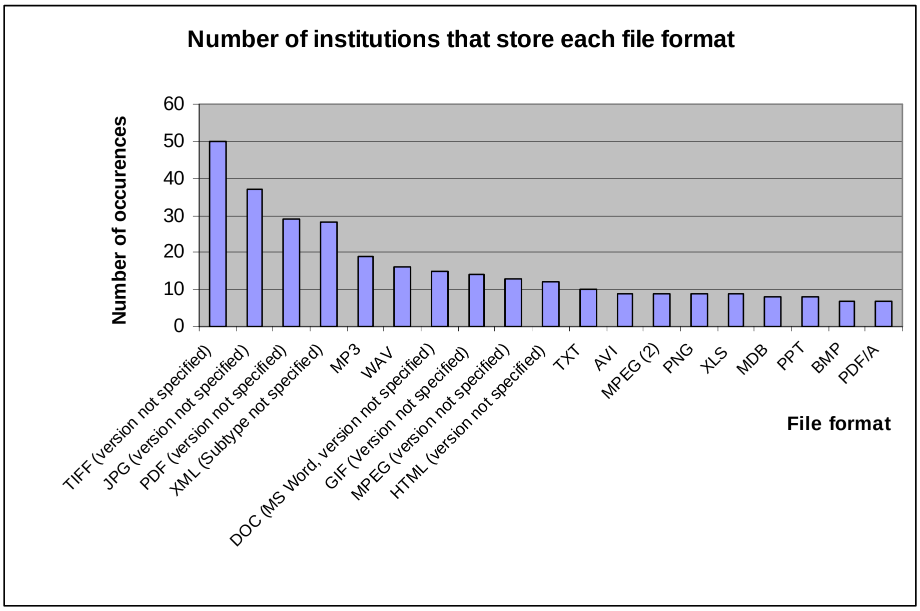

Figure 1. Number of institutions that archive each file format.

There are several file formats that are archived by many institutions; two are found in 50% of the institutions. These two are:

- TIFF (in 50 of 76, or 66% of the institutions),
- JPEG (in 37 of 65, or 49% of the institutions).

Other file formats that occur in a relatively large amount of institutions are:

- PDF (38%),
- XML (37%),
- MP3 (25%),
- WAV (21%),
- DOC (20%),
- GIF (18%),
- MPEG (17%),
- HTML (16%).

Within this report, libraries, archives and museums are the main categories of institutions. All three of them archive mostly TIFF, JPG and PDF. However, further down the list of archived file formats within each category, differences become apparent, like the fact that archives hardly archive the file format MP3, while libraries and museums do. This is a rather significant observation as hardly used formats may turn out to be very important to a certain type of institution. For example, such as the DOC format, that turns out to be one of the most used file formats in archives.

2.5 Overview

It has been difficult to gather information about the archived file formats by cultural heritage institutions worldwide. The reason for this seems to be that there has not been much research into this subject yet. Also, institutions do not seem to know exactly which file formats they archive in general for the long term. Still, it was possible to produce a substantial list to support further research into possible gaps between file formats and preservation action tools.

After extending the list of file formats both in size as in content several times, the inventory of file formats didn't change significantly anymore. This confirms that the top file formats found in the previous iteration were the most archived file formats in cultural and scientific institutions. A few new file formats that have been found have been added to the inventory. This might indicate that there are many file formats only used by a small selection of institutions. These file formats must be looked at, but treated differently than the main file formats.

When compared to existing studies, the file format inventory is mostly confirmed. In a few cases the picture painted by these studies is different, indicating that there might be difference based on location or type of institution. The differences are small however, and deal more with the ranking of file formats than the occurrence. The added value of the gap analysis is that this is the first study that is focussed on the file formats and available preservation action tools, whereas these existing studies have information about file formats as a side question to support another goal. Also, the types and geographical locations of the institutions surveyed are much broader in the gap analysis.

The file format inventory now contains 137 file formats, submitted by 76 respondents. The realisation of this inventory is of vital value for establishing gaps in tool provision which is the ultimate purpose of this research.

In the following chapters some case studies are being discussed as the above showed that only a few file formats are archived by many institutions. This might indicate that a very specific file format could be very important for a certain type of institution which in turn might mean very specific preservation action tools are needed.

Also, an inventory of preservation action tools must be created to compare with the file format registry.

3. Preservation action tools

3.1 Introduction

Preservation action tools have been defined as “A software program that performs a specific action on a digital object to ensure the continued accessibility of this digital object”. Broadly, these tools can be divided into two mains categories: tools that change the object, and tools that change the environment in which the object is accessed.

Tools for objects are usually classified as migration tools, and fall under work package PA/4 in Planets. This work package has compiled a list of existing migration tools. Tools for environments are classified as emulation tools, and fall under PA/5.

3.2 Methodology

The work packages that cover these tools, PA/4 and PA/5 have set up lists of migration and emulation tools. This list contains migration and emulation tools that are known and have been used by the Planets partners who made this list and is by no means complete. However, it is a good indication of commonly used tools. The tools on this list will be the first tools that are wrapped as a service and made available in the Planets Framework.

Other lists and overviews of migration and emulations are published on various websites. Google Directories has a list in the Data-Format > Conversion > Software category[3], and the website 4convert (although it seems dormant) offers an overview of specific migration actions and the tools that can perform them[4]. A simple search on the internet reveals even more tools.

Therefore it is important to define what kind of migration tools should be the focus of this gap analysis. As the intended audience is interested in long-term preservation, the tools should be suitable for that. Within this report there will be no tests undertaken to determine the usefulness of certain tools for long-term preservation; these tests will eventually be carried out on the Testbed.

The initial gap analysis will therefore start out with the list of tools compiled by the Planets partners, and extend this for the five most used file formats.

3.3 Migration tools

The list of migration tools (Appendix B) known and used by Planets Partners contains 57 tools. These tools all use one of three ways to interact with the user: a graphical user interface (GUI), a command line and availability as an online service. Each interface has its own advantages and disadvantages which will influence the gap analysis. For example, online tools maybe impossible to wrap and plug into an existing storage and maintenance system, and can therefore be useless to an organization in search of a tool. There is also a wide range of licenses under which these tools are available. This will also influence the usability of each tool.

3.4 Emulation tools

When looking at emulation tools, we are specifically looking at hardware emulators. For emulation these tools, the list is much shorter, but not any less complicated. Because emulation emulates “just” the hardware, it is up to the organization that uses the tool to supply the necessary operating system and software to make object accessible. If all software (operating systems and rendering programs) is available, a long list of file formats can be accessed.

[3] http://www.google.com/Top/Computers/Data_Formats/Conversion/Software/

[4] http://www.4convert.com/

Because the usability of an emulation tool for a specific digital preservation solution depends heavily on the software archive of each institution or on available free alternatives, it is very difficult to give a definitive answer to which file formats can be accessed through each emulator. Only general statements can be made about a possibility of each emulator being able to access a given file format if a predetermined set of other software is available.

3.5 Planets Core Registry

One of the products in Planets is the Planets Core Registry (PCR). The PCR is a follow-up version to Pronom, the file format registry developed and maintained by The National Archives. The PCR contains detailed information about file formats, software and hardware. Each entry in the PCR is identifiable by a unique ID (PUID).

A record of a file format contains information about the software related to the file format. This can be software used to render or create an object in that file format, but can also be a preservation action tool that can either migrate or access the object. On the software record all kinds of information is recorded, including which file formats it can interact with, as well as which actions it can perform.

The part of the registry that is most interesting to the Gap Analysis is the possibility to add Pathways. A Pathway describes a migration, emulation or characterization action. This uses an input file format, one or more tools and an output file format (in the case of a migration action).

The PCR will be accessible for users through an online user interface. Web services will be available for integration of the PCR in other applications. The PCR will be integrated into the Planets Framework; Testbed results will be stored in the PCR, and Plato will use the PCR as a source of information.

Using the PCR and the Pathways that are described it should be possible to identify any existing gaps. In theory, if there are no pathways associated with a file format, a gap exists.

Unfortunately, the PCR is as yet unavailable for testing this theory for two reasons. The current version (PCR 2.0) has only recently been released in a limited release. Also, the current content is taken directly from Pronom, and is continuously updated with information about preservation action tools and pathways. This means that the exploration of the use of the PCR for the gap analysis is an exercise in theory for now.

3.6 Migration tools

In the future potentially all migration tools will be described in the PCR. For all described tools it will be known which file formats are the intended input and output file formats for each tool. These tools will be linked directly to the relevant file format entities. This enables a user to instantly see if a file format has any tools associated with it, and what the properties of these tools are.

3.7 Emulation tools

In theory the registration of emulation tools will happen in the same manner as migration tools. However, due to the nature of these tools they cannot be linked directly to file formats. An emulation tool will have an associated hardware configuration. A software package, representing an operating system, will be linked to this hardware entity, creating a technical environment. Regular software programs, also registered as software packages, will be linked to the operating systems. File formats are linked to these programs.

Currently, there is no automatic way to discover if a file format can be accessed with the use of an emulator. The reverse is possible; a record for an emulator will enable the user to discover all programs that can run on the emulator.

3.8 Pathways

Pathways will enable the user to quickly get an overview of preservation actions that are available for a file format, whether they are migration or emulation actions. The user (human or system) can search the database for pathways by specifying an input file format and an optional output file format.

Of course, the full potential of the PCR will only be released if sufficient complete data is entered into the database. This will take some time, especially in the case of software. For now, pathways must be manually entered by an administrator. This means that entering an emulation tool with all possible pathways is a daunting task. In the future this may be automated. However, it may be possible to search the underlying database directly for file formats without an access possibility through an emulation tool.

3.9 Conclusion

Making concrete statements about the existence of gaps in tool provision depends on the availability of complete accessible information. For now, this information is scattered. An effort is made to provide an overview of all relevant information in the Planets Core Registry. Unfortunately, the PCR is still in development, making the usefulness of the PCR for this report purely theoretical. However, it is certain that when the PCR (or any other digital preservation registry) contains enough information about file formats, tools, programs etc. it will be an extremely useful tool to identify any and all gaps in tool provision.

Even though a full comparison between the compiled list of file formats and all existing preservation action tools in the world is impossible right now, and may never be possible, there are some comparisons that can be made.

For instance, the ten most occurring file formats on the file format inventory can be compared to the list of available tools, to determine if those file formats have tools available for them. Looking at the list, we find this to be the case, with the exception of one file format, XML. There are no tools on the list specifically designed to migrate XML files to another file format. The reason for this is that XML is probably the designation file format in many migration and normalization actions because it is an open and easily accessible file format. This is visible in that XML is a destination file format for many migration tools.

All other file formats have one or more tools on the list that can take that format as an input file format and migrate it to another file format. In the strictest sense, this means that there are no gaps in tool provision where those file formats are concerned.

However, as we will see in the next chapter the search for gaps cannot be limited to availability of tools for most occurring file formats. There are many specialized formats out there that need support from specialized PA tools.

4. Gap analysis

4.1 Introduction

To answer the question “Are there gaps in tool provision” first the term gap needs to be defined. In the question, a gap can be found if there are no tools available for a given file format. However, this might not be the whole answer, as each organization has its own requirements for a preservation action tool.

If a tool is only available in a Unix environment, and the organization can only use Windows tools, that organization experiences a gap. If the available tool cannot be licensed for use in the organization, there is a gap. If a tool is available for a certain input file format (for example DOC), but it cannot output the desired output file format, there is a gap. If a tool is available, but it cannot preserve certain properties of the object, there is a gap.

Therefore there can never be one conclusion to the search for gaps in tool provision, because the gap is defined differently by each and every organization. However, this does not mean there are not some conclusions that can be drawn from the previously presented research. In this chapter we will look at emulation and migration tools in general, the Plato tool and three case studies of some niche file formats.

4.2 Emulation tools

There are two main problems with defining gaps in the availability of emulation tools for digital preservation. The first problem is that an emulator cannot be used on its own. The organization always needs software in its own collection to enable access to the object. This means that the organization must own the operating system and software and the right to use it. It is impossible to determine this availability, which means that no general statements can be made about the usefulness of a certain emulator. The only thing that can be said is that an emulator can access a certain file format in theory, if the organization owns all required additional software.

The second problem is that since emulators enable an organization to run a legacy operating system, in theory the number of file formats that can be opened is nearly endless. Like described before, it is nearly impossible to register all file formats that can be accessed by an emulator. Therefore, using emulation, an organization can in theory open nearly every file format, provided they have the required software available to them.

4.3 Migration tools

To define a basic gap (is there a tool available) is relatively easy. As long as tools are registered in the PCR, it can easily be determined which file formats have no tool available. This does require the information in the PCR to be complete and extensive.

Some tools are problematic however. There are tools in existence that can take almost all file formats that can be opened as input. Examples are tools that can 'print' to PDF or tools that wrap file formats to XML. The complete list of file formats these tools can work on is unavailable because it is constantly being extended, as any file format that can printed can be converted.

Also, the availability of a tool says nothing about how well specific requirements are fulfilled by a tool. Not until a tool has been tested in the Planets Testbed or any other test environment can anything be said about the usefulness of a tool in a specific digital preservation situation.

4.4 Plato

Within Planets, the preservation planning tool Plato is being developed. This tool enables organisation to plan their preservation action for a specific collection of digital objects. To do this, organisations are required to fill in an extensive list of requirements. Together, all the requirements provide the organisation with a preservation plan for this specific collection. Based on this plan, Plato chooses the best available preservation action tool.

Plato executes a search for these tools in the Planets Core Registry. In effect, if the Plato tool cannot provide a suitable preservation action tool that can perform the desired action while complying with all policy requirements, this means that there is a gap in the availability of those tools. Of course, this can only be said if the PCR is complete and up to date with regards to preservation action tool information.

4.5 Case studies of individual file formats

The analysis of the file format inventory shows that only a few file formats are archived by many institutions. However, that does not mean that the other file formats are not important and should be ignored. One such file format could be very important for a certain type of institution which does not find an alternative in one of the "bigger" file formats. This leads to the assumption that if a file format is important for an institution or a small but specialized group of institutions, and there are problems with that file format (i.e. release of a new version), that institution or group of institutions will provide their own solution to these problems. This assumption is tested in this chapter with three case studies into file formats that are not wide spread, but important to the institutions that use and store them.

4.5.1 Sheet music file formats

4.5.1.1 The file formats

In the world of digital western sheet music there are several major file formats. All but one of those are proprietary file formats created and used by commercial software, one format is an open source initiative.

4.5.1.2 Use of sheet music file formats

To write sheet music using a computer, scorewriting software is used. This can be compared to the use of a word processor for the writing of text. A scorewriter allows the user to input, edit, print and exchange music in the form of sheet music.

4.5.1.3 Sheet music file formats and digital preservation

The problem with scorewriting file formats and digital preservation is that there are several commercial players in the field who each have their own proprietary format, and have no need for a shared standard format. There is no clear main player in the market (like Microsoft Office Word is for word processing for example), so there is no clear format "to bet on" for the future. Ideally, each file should be saved in all possible formats to insure that the file can be opened by future software.

Each scorewriter has its own propriety format. Import methods of files from competitors are limited or non existent. Other input methods are manual, MIDI import (music can be imported while it's being played) and music OCR. Export methods from each program are also limited to the format tied to the program, or PDF which does not offer the same possibilities to its users.[5] All import and export methods are unable to capture all the details within the files.

Also, the commercial scorewriters publish regular updates to their software. The use of these new releases requires conversion of files saved by previous releases, which makes digital preservation of these files even harder.

MusicXML, the open source sheet music file format is a good alternative. It is open, and thus can be used by anyone.[6] Its use is implemented in most commercial score writers (import and export functionality), and therefore the format has a big potential user base. There are no dedicated migration tools available. Development of the format is done by a commercial corporation;

[5] Michael Good, MusicXML: An Internet-Friendly Format for Sheet Music, <http://www.idealliance.org/papers/xml2001/papers/html/03-04-05.html>, accessed on August 21, 2008

[6] Recordare LLC, < http://www.recordare.com/xml.html>, accessed on September 18, 2008

however, the W3C has also released an XML schema definition (XSD) for MusicXML. This makes the file format even more regulated.

This all leads to the conclusion that MusicXML might be the best solution for the preservation of music at present.

4.5.1.4 Available solutions for problems with sheet music file formats

The main problem with sheet music file formats is the exchange of information. This is difficult due to the limited import and export features of each program. The solution is not easy; each software company must adapt its own products to provide the import and export functionality.

MusicXML might be a good alternative, however, each software provider must provide integration with MusicXML and so the viability of this solution is dependant on each individual software company. Recordare, the company that developed and maintains MusicXML maintains a list of migration utilities and status of integration of MusicXML with each software package.

4.5.1.5 Summary

The world of sheet music software and its formats is very fragmented. There are many file formats and not one is the most popular. There is not one organisation that develops one or more file formats that are the standard; it is many different developing institutions that develop mostly proprietary file formats. Because of this there is little or no cooperation between the file formats and the software packages, making digital preservation a hard task.

4.5.2 FITS

4.5.2.1 The file format

FITS stands for “Flexible Image Transport System”. It is a format used to store astronomical data and is the standard format in use for this sort of information. FITS is endorsed by organizations such as NASA and the International Astronomical Union (IAU). The development of FITS is managed by the IAU FITS Working group, which in turn is supported by four regional committees (Australia/New Zealand, Europe, Japan and North America).[7]

In the late 1970s the astronomical field found that with all the new and emerging techniques in astronomy, the number of digital images they had to work with increased dramatically. To work with or compare data, astronomers wished to transport data to their own institution. The problem that arose was that each institution traditionally had developed software systems and data formats for their own use. Exchange of data between institutions was impossible due to incompatible systems. The solution to this problem was the development of a format, FITS, which would be used to transport data from one institute to another. Each institute only needed to write a translation program that could convert their own data format to FITS and vice versa.[8] Today FITS is the standard file format used by astronomers world wide.

4.5.2.2 Use of FITS

In the file format inventory, FITS is found just once, at the Wide Field Astronomy Unit, School of Physics, University of Edinburgh in Scotland. This is the only institution in the survey that works specifically with astronomical data.

Despite only finding FITS once in the file format inventory, the use of FITS is widespread in all fields of astronomy, in all countries. The International Astronomical Union has designated FITS as the standard for astronomical data, and as such, many, if not all, of their members use FITS.

4.5.2.3 FITS and digital preservation

Because FITS is a scientific format, there are only a few programs that can deal with FITS as a file format for manipulation (not just transport). If FITS is used for transport of data, each institution must have its own interpretation program. This means that for access there only a few programs that need to be updated when a new version is released.

[7] IAU-FWG, <http://fits.gsfc.nasa.gov/iaufwg/iaufwg.html>, accessed on September 9 2008

[8] D.C. Wells, E.W. Greisen and R.H. Harten, “FITS: A Flexible Image Transport System”, *Astronomy & Astrophysics Supplement Series* 44 (1981), 363-370

Each new version of FITS offers new features, but existing aspects are rarely depreciated. If an aspect is depreciated, this is usually an aspect whose use was restricted or discouraged before.

If an institution wants to benefit from all available new features, it must update its programs with each new version. When a program is not updated, it will be able to interpret a new version, however new features will be unavailable.

New versions are released very sporadically, with demand from the user community being the main reason. This makes the adoption rate of a new format, and the necessary software updates, much higher. However, because the software for the use of FITS is tailor made by each institution that uses it, FITS is still a risky format for digital preservation.

4.5.2.4 Available solutions for FITS problems

Because of the reasons described above, there are next to no problems with the accessibility of FITS as long as the information about each version of the standard is available. The only main problem that threatened the use of FITS was the well known year 2000 date problem. This problem was discovered early on, and a solution was provided in 1997 with the adoption of a date format that was Y2K compliant.

The FITS support office at the NASA/GSFC offers documentation that helps institutions and software providers to solve their own problems. Detailed description of the format, and changes in comparison to previous releases, are provided, as well as packages for numerous programming languages to aid programmers in making their program work with FITS. The support office also offers information about programs that work with FITS, such as validators, viewers and editors.

4.5.2.5 Summary

The FITS format is used by a relatively small group of institutions world wide. However, it is used almost exclusively in one specific sector, that of astronomy. The development of FITS is managed by a central working group with support from the end-users of the file format. Any changes to the standard are communicated through this working group, which also offers resources to be able to cope with these changes.

This leads to a situation where any issues with the format or its software are identified and solved within the community. However, the format is still not ideal, because all solutions are tailor made by each institution.

4.5.3 DAISY

4.5.3.1 The file format

DAISY stands for Digital Accessible Information System and is a file format standard that is used to make talking books available to users with reading disabilities.[9] The file format standard is open source. A DAISY Book is a set of files that contains:

- One or more audio files with human narration of the text (MPEG 4, MP3, WAV);
- A marked-up file containing the text (optional) (XML);
- A synchronization file that links the marks in the text with time points in the audio file(s) (SMIL);
- A navigation control file that enables navigation (NCX).

The DAISY standard (ANSI/NISO Z39.86) has been developed by the DAISY consortium (an organization with over 70 full or associate members from all over the world), The National Library Service for the Blind and Physically Handicapped (part of the Library of Congress), and several other organizations in North America. The current version is DAISY 3.

DAISY is an enhanced audiobook that offers better possibilities for navigation. This would enable an user with a reading disability to navigate through an encyclopedia in audio format.

Digital talking books (DTB) in the DAISY format are typically available on CD-Rom. One such CD-Rom can contain up to 50 hours of narration.[10] This CD-Rom can then be played by a special

[9] DAISY Consortium, *About us*, < http://www.daisy.org/about_us/index.shtml>, accessed on August 11, 2008.

player or with a computer. There are several manufacturers of these DAISY players who are also member of the DAISY consortium. DAISY DTB's can be played on a computer by using special software. Without this software, individual files in the DAISY Book might be accessible (for example the MP3 of the audio) but not in the way that was intended by the creator of the file for visually impaired users. Of course, with the advance of the internet, DAISY DTB's are also available as downloads to users.

4.5.3.2 Use of DAISY

In the file format survey of the previous iteration, one institution indicated they archive files in DAISY format for the long term. However, there are a few institutions in each country that use and archive DAISY files as part of a national service to those with reading disabilities. Examples are *Loket aangepast lezen* in the Netherlands, *Danmarks Blindebibliotek* in Denmark and *Talboks- och punktskriftsbiblioteket* in Sweden.

This means that even though the inventory of archived file formats showed that very few (mainstream) institutions archive DAISY files, the format is important to a large and international group of people.

4.5.3.3 DAISY and digital preservation

A DAISY DTB can be treated as a regular file. Thus it has the same problems concerning digital preservation as any other file. The DAISY standard evolves, and is updated. A new version offers more possibilities, old hard- and software must be updated to be able to handle this new functionality.

Also, if new operating systems are released, new software might be needed to be able to use DAISY files. This new software should ideally be able to handle both the most recent version of the DAISY format, as well as all versions that came before that.

4.5.3.4 Available solutions for DAISY problems

Possible problems with DAISY arise with the adoption of a new version of the file format. The DAISY consortium offers help in such cases.

The DAISY consortium offers a roadmap for the implementation of the newest DAISY standard (upgrade from DAISY 2.02 to DAISY 3) in which they give tips on how to optimize DAISY 2.02 files. Also, the consortium offers several tools to help with the implementation of DAISY 3, such as a validator for DAISY 2.02 and an automated migration tool to convert files from DAISY 2.02 to DAISY 3.0.[11]

The DAISY consortium also developed a data model to help producers with the development of soft- and hardware tools for DAISY playback. However, development of said tools is not controlled or regulated by the consortium, so there is no guarantee for backward compatible playback tools.

4.5.3.5 Summary

The development of DAISY is comparable with the development of FITS, as it is also managed by a central commission. In the case of DAISY the consortium is made up of content providers, not end users. This is slightly different than the FITS working group, however, due to the difference between end-user groups, this works better in the case of DAISY.

As with FITS, problems that arise are identified and solved within the consortium.

4.5.4 Conclusion

The three case studies presented above show three different situations. For sheet music file formats, there is no central coordination. With FITS development is coordinated by a group of end users. For DAISY, development is coordinated by service providers.

FITS and DAISY provide little to no specific problems for digital preservation. Their standards are published openly, and old standards are still available. When problems arise due to the development of a new version of the format, new operating systems and new hardware, these

[10] Talboks- och punktskriftsbiblioteket, < http://www.tpb.se/english/talking_books/general_daisy_information/>, accessed on August 18 2008

[11] DAISY Consortium, *Road Map to Implementation of DAISY*, <http://www.daisy.org/z3986/DAISY.Consortium.Roadmap.2005-09.html>, accessed on August 18 2008

problems are picked up by the central organization that works on the file format. Solutions are developed by users and service providers, and information is available centrally to facilitate these solutions.

With sheet music formats problems are much bigger because there is no central organization that can put pressure on each software producer to facilitate the possibility of exchange of information. This means that problems are solved only when this is wanted by the software producers (i.e. it offers an opportunity for a new sale), not when this is needed by users. Solutions are therefore only developed when someone with the specific knowledge and access feels the need, not when the user community needs it. This is not only a problem with regards to accessibility of information now, but also later. Without pressure from an organized community there is chance that software producers might not help with the problems their files might occur when archived for the long term.

The development of a file format, or file format type, which is regulated centrally, provides fewer problems for digital preservation, because solutions are developed within the user or developer communities of those file formats. When the development is not regulated, solutions are sporadically developed. For digital preservation, these file formats should be kept under watch.

4.6 Conclusion

Without specific limitations within the question, the answer to question “is there a gap” is almost certainly “no”. The question is not specific enough. However, no conclusions can be drawn from this. Even though there are most likely no gaps, there are still preservation action tools that are missing. It is very likely that when an institution places specific demands on a preservation action tool, they may find that the currently existing tools are lacking. The problem is that this gap will not be defined until a specific preservation action question is asked.

Examples show that, for specific institutions, there might be a gap, even though this study might have found none. It is impossible to define all demands an institution might have for a PA tool, and analyze all these demands to find out if there are tools available to meet these demands. Therefore it is impossible to say if there are gaps in the availability of preservation action tools.

When Plato is used by institutions for preservation planning, more concrete statements can be made about the availability of desired tools. Every time Plato cannot find a suitable tool for the requirements provided by an institution, there is a gap in the availability of tools. These gaps can be inventoried to guide the development of new preservation action tools.

It is important not to overlook the file formats that might not occur more than once or twice in the file format inventory. Each file format might serve a small but dependable market, as shown in the three case studies. If this is the case, digital preservation problems might be taken care of by the market or the file format developers. This is the case with FITS and DAISY, both file formats are open standards that are regulated by a central non-profit consortium. However, when the file formats are developed and maintained by unregulated for-profit, hardly any digital preservation solutions are developed, placing the file format at risk.

5. Conclusion

An extensive survey into the file formats archived by cultural heritage and scientific institutions was undertaken. The inventory now contains 137 file formats, submitted by 76 respondents. The list of file formats shows that only a few file formats are archived by many institutions. This means that the availability of tools for these file formats is very important. However, a very specific file format could be very important for a certain type of institution which in turn might mean very specific preservation action tools are needed.

Complete information about preservation action tools will be gathered in the Planets Core Registry. Unfortunately, the PCR is still in development. The information in the PCR is not yet complete, making the usefulness of the PCR for this report purely theoretical. However, it is certain that when the PCR (or any other digital preservation registry) contains enough information about file formats, tools, programs etc. it will be an extremely useful tool to identify any and all gaps in tool provision.

Even though a full comparison between the compiled list of file formats and all existing preservation action tools in the world is impossible right now, and may never be possible, there are some comparisons that can be made.

An inventory of migration tools gathered at Planets partners is made and contains 57 tools. These tools can convert the ten most used file formats into another file format, except for XML. The theory is that XML is often the preservation format of choice; many tools convert to XML, none convert from XML.

A straight answer to the question “Are there gaps in tool provision” is “no”. However, a gap is defined differently by each and every organization because of the specific requirements that each organization has. In the case of specific functionality or quality demands, it is almost certain that gaps in tool provision will remain to be found. If these organizations use Plato for their preservation planning, any time Plato cannot find a suitable tool, there is a gap in tool provision. These gaps cannot be found by only looking at the compiled lists of file formats and preservation action tools.

However, it is important not to overlook the file formats that only occur once or twice in our file format list, because even though these file formats might not be used much by mainstream organisations, they might be important for a niche group. Three of these file formats were explored in case studies, and it was found that when an open consortium of users or developers work together, there is a good chance that any digital preservation issues will be overcome by the users and/or developers. If development of the file format and associated software is fragmented, digital preservation issues will arise and will be much more difficult to solve.

The world of digital objects and digital preservation is constantly evolving. New file formats are adopted, new tools are developed. More information about file formats and tools will be gathered and made available. For these reason the gap analysis should be redone every two to three years to take full advantage of the new information that becomes available.

The results of the analysis not only give important insight in the status of digital preservation, but also in the status of digitization. Even more important a regular gap analysis in tool provision can be used in a justification and indication for new research and/or development of new tools.

6. Appendices

Appendix A List of found file formats

The list is sorted by number of occurrences.

File type	PUID	Number
TIFF (version not specified)	fmt/7 - fmt/8 - fmt/9 - fmt/11	50
JPG (version not specified)	fmt/41 - fmt/42 - fmt/43 - fmt/44	37
PDF (version not specified)	ftm/14 - fmt/15 - fmt/16 - fmt/17 - fmt/18 - fmt/19 - fmt/20	29
XML (Subtype not specified)	fmt/101	28
MP3	fmt/134	19
WAV	x-fmt/389 - x-fmt/396 - x-fmt/397	16
DOC (version not specified)	fmt/37 - fmt/38 - fmt/39 - fmt/40 - x-fmt/2 - x-fmt/129	15
GIF (version not specified)	fmt/3 - fmt/4	14
MPEG (version not specified)	x-fmt/385 - x-fmt/386	13
HTML (version not specified)	fmt/96 - fmt/97 - fmt/98 - fmt/99 - fmt/100 - fmt/102 - fmt/103	12
TXT	x-fmt/14 - x-fmt/15 - x-fmt/130 - x-fmt/111 - x-fmt/110	10
AVI	fmt/5	9
MPEG (2)	x-fmt/387	9
PNG	fmt/11 - fmt/12 - fmt/13	9
XLS	fmt/55 - fmt/56 - fmt/57 - fmt/59 - fmt/60 - fmt/61 - fmt/62	9
MDB	x-fmt/66 - x-fmt/238 - x-fmt/239 - x-fmt/240 - x-fmt/241	8
PPT	fmt/125 - fmt/126 - x-fmt/88	8
BMP	x-fmt/25 - x-fmt/270 - fmt/114 - fmt/115 - fmt/116 - fmt/117 - fmt/118 - fmt/119	7
PDF/A	fmt/95	7
CSS	x-fmt/224 - x-fmt/145	5
MOV	x-fmt/384	5
MPEG (1)	x-fmt/385	5
PSD	x-fmt/92	5
RA	x-fmt/278	5
RTF	fmt/45 - fmt/46 - fmt/47 - fmt/48 - fmt/49 - fmt/50 - fmt/51 - fmt/52 - fmt/53	5
SWF	fmt/104 - fmt/105 - fmt/106 - fmt/107 - fmt/108 - fmt/109 - fmt/110	5
WAV (Broadcast Wave, BWF)	fmt/1 - fmt/2	5
ZIP	x-fmt/263	5
ASCII	x-fmt/22 - x-fmt/283	4
CSV	x-fmt/19	4
MP4		4

File type	PUID	Number
ODF		4
WMF		4
DWG	fmt/21 - fmt/22 - fmt/23 - fmt/24 - fmt/25 - fmt/26 - fmt/27 - fmt/28 - fmt/29 - fmt/30 - fmt/31 - fmt/32 - fmt/33 - fmt/34 - fmt/35	3
EXE	x-fmt/409 - x-fmt/410 - x-fmt/411	3
JPEG2000	x-fmt/392	3
PS	x-fmt/91 - x-fmt/406 - x-fmt/4-7 - x-fmt/408	3
QXD	x-fmt/182	3
SPSS (portable)		3
TIFF 6.0	fmt/11	3
WMA		3
WordPerfect (version not specified)	x-fmt/44 - x-fmt/203 - x-fmt/393 - x-fmt/394	3
XHTML	fmt/102 - fmt/103	3
ARC		2
DBF	x-fmt/8 - x-fmt/9 - x-fmt/10 - x-fmt/271 - x-fmt/272 - x-fmt/380	2
DBF	x-fmt/8 - x-fmt/9 - x-fmt/10 - x-fmt/271 - x-fmt/272 - x-fmt/380	2
DPX		2
EPS	fmt/122 - fmt/123 - fmt/124	2
FLASH		2
JS	x-fmt/423	2
MPP	fmt/243 - fmt/244 - fmt/245 - fmt/246 - fmt/247	2
MSG		2
NEF		2
NSF		2
PCD (PhotoCD)		2
PPS	x-fmt/87	2
SGML	x-fmt/196	2
SVG	fmt/91 - fmt/93	2
TAR	x-fmt/266	2
WordStar	x-fmt/370 - x-fmt/260 - x-fmt/205 - x-fmt/236 - x-fmt/237 - x-fmt/261 - x-fmt/206 - x-fmt/262	2
AIFF	x-fmt/135	1
ALTO		1
ARW		1
ASF		1
ASP	x-fmt/138	1
AU	x-fmt/139	1

File type	PUID	Number
BIN		1
CDR	x-fmt/29 - x-fmt/291 - x-fmt/292 - x-fmt/374 - x-fmt/375 - x-fmt/378 - x-fmt/379	1
DCR		1
DOC (MS Word 97-2002 Document)	fmt/40	1
DOT	x-fmt/45	1
DTB		1
DV		1
DVCAM		1
DVCPRO (version unspecified)		1
DVCPRO25		1
DVCPRO50		1
ENL		1
ESRI-shape	x-fmt/235	1
FITS		1
FLAC		1
FM	x-fmt/302	1
FP3	x-fmt/318	1
FP5	x-fmt/319	1
FP7		1
Freehand		1
GIF89a	fmt/4	1
GIS		1
GM		1
GML	x-fmt/227	1
GZIP	x-fmt/266	1
INDD		1
INF		1
ING		1
INI		1
INX		1
JAR		1
JSP	x-fmt/160	1
LOG	x-fmt/62	1
Mini-disc RAW		1
MJ2		1
MP2		1

File type	PUID	Number
MrSID		1
MXF		1
NB		1
newsML		1
ODT	fmt/136	1
OEBPS		1
OGG		1
PDF 1.1	fmt/15	1
PDF 1.2	fmt/16	1
PDF 1.3	fmt/17	1
PDF 1.4	fmt/18	1
PDF 1.5	fmt/19	1
PHP	x-fmt/169	1
PICT		1
POT	x-fmt/84	1
PSP	x-fmt/233 - x-fmt/234 - x-fmt/297 - x-fmt/298 - x-fmt/376 - x-fmt/377	1
PST	x-fmt/248 - x-fmt/249 - x-fmt/250 - x-fmt/251	1
RV		1
SD2		1
SMIL		1
SPIFF		1
SQL		1
STR		1
SunAU		1
TIFF 5.0	fmt/9	1
TXT (IBM DisplayWrite 2 & 3)	x-fmt/288 - x-fmt/289	1
VSD	x-fmt/113 - x-fmt/258 - x-fmt/259	1
WK1	x-fmt/114	1
WK3	x-fmt/115 - x-fmt/116	1
WK4	x-fmt/116 - x-fmt/117	1
WK5	x-fmt/212	1
WordPerfect 5.1	x-fmt/394	1
WRL	fmt/93 - fmt/94	1
XHTML 1.0	fmt/102	1
XLA	x-fmt/123 - x-fmt/124	1
XLS (MS Excel 97-2002 Workbook)	fmt/61 - fmt/62	1

Appendix B List of migration tools

Name	Creator
2007 Microsoft Office Add-in: Microsoft Save As PDF or XPS	Microsoft
AbiWord	Open Source
ACDSee	ACD Systems
AviDemux	Open Source
BullZip Printer	BullZip
ConServR	Lincoln & Co
Cumulus	Canto
CZ-Doc2Pdf 2.0	ConvertZone Software co., ltd.
Dia	Open Source
DioscuriArjMigration (Dioscuri)	Open Source (KB-NL, NANETH, PLANETS)
DioscuriPnmToPngMigration (Dioscuri)	Open Source (KB-NL, NANETH, PLANETS)
Document2PDF Pilot 1.10	Two Pilots
dvips	Open Source
EscapeE	RedTitan
Express Server	Adlib software
Extractor (XcdlMigrate)	Planets
ffmpeg	Open Source
Ghostscript	Open Source
Gimp	Open source
Graphic Converter	Lemkesoft
Graphics Magick	Open Source
ImageMagick	ImageMagick Studio LLC
InkScape	Open Source
Jasper19	Open Source
Java-SE	Sun Microsystems
JJ2000	JJ2000 partners (Canon, Ericson, Swiss Federal Institute of Technology)
JTidy	Open Source
Media Convert	Media Convert
MEncoder	MPlayer Team
Microsoft Office Compatibility Pack for Word, Excel, and PowerPoint 2007 File Formats	Microsoft
MsgText	Enterag
MyMorph	Lister Hill National Center for Biomedical Communications - National Library of Medicine
NetPBM	Open Source
Office File Converter Pack	Microsoft
OpenJpeg	Open Source
PDF online	BCL Technologies
PDF Version Converter	NicePDFSofware Inc
PDF/A Converter	May computer
Pdf2PdfAMayComputer	May computer
Pdf2Ps	Unix shell script
PdfBox	Open Source
PhotoShop	Adobe
Print2PDF SE 6	Software602, Inc
ps2pdf	Open Source
Purepage SDK	Inzone Software Limited
SanselanMigrate	Apache Foundation

Name	Creator
SIARD	Swiss Federal Archives
Silentprint	FunAsset
SoX	Open Source
Transformation Server	Stellent
TRYNT HTML Converter Web Service	Trynt Heavy Technologies
Universal Document Converter	fCoder Group, Inc.
UvcMigrate	IBM
VERS	Public Record Office Victoria (PROV)
VisualIntegrity	Visual Integrity Technologies
XENA	National Archives Australia
Zamzar	Zamzar